Volume 47 Number 3 September 2023 ISSN 0350-5596

Informatica

An International Journal of Computing and Informatics

Special Issue:

SoICT 2022

Guest Editors:

Huynh Thi Thanh Binh,
Ichiro Ide

1977

Editorial Boards

Informatica is a journal primarily covering intelligent systems in the European computer science, informatics and cognitive community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor from the Editorial Board can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the list of referees. Each paper bears the name of the editor who appointed the referees. Each editor can propose new members for the Editorial Board or referees. Editors and referees inactive for a longer period can be automatically replaced. Changes in the Editorial Board are confirmed by the Executive Editors.

The coordination necessary is made through the Executive Editors who examine the reviews, sort the accepted articles and maintain appropriate international distribution. The Executive Board is appointed by the Society Informatika. Informatica is partially supported by the Slovenian Ministry of Higher Education, Science and Technology.

Each author is guaranteed to receive the reviews of his article. When accepted, publication in Informatica is guaranteed in less than one year after the Executive Editors receive the corrected version of the article.

Guest Editorial Preface
Information and Communication Technology

Since 2010, the Symposium on Information and Communication Technology—SoICT has been organized annually. The symposium provides an academic forum for researchers to share their latest research findings and to identify future challenges in computer science. The best papers from SoICT 2015, SoICT 2016, SoICT 2017, and SoICT 2019 have been extended and published in the Special issues "SoICT 2015", "SoICT 2016", "SoICT 2017", and "SoICT 2019" of the Informatica Journal, Vol.40, No.2 (2016), Vol. 41, No. 2 (2017), Vol. 42, No. 3 (2018), and Vol. 44, No 2 (2020), respectively.

In 2022, SoICT was held in the scenic Ha Long Bay, Vietnam, from December 1–3. The symposium covered four major areas of research including Artificial Intelligence and Big Data, Information Networks and Communication Systems, Human-Computer Interaction, and Software Engineering and Applied Computing.

Among 102 submissions from 14 countries, 42 papers were accepted for oral presentation at SoICT 2022 and 20 papers for posters. Among them, the following six papers were carefully selected, after further extension and additional reviews, for inclusion in this special issue.

The first paper, "Lightweight Multi-Objective and Many-Objective Problem Formulations for Evolutionary Neural Architecture Search with the Training-Free Performance Metric Synaptic Flow" by An Vo, Tan Ngoc Pham, Van Bich Nguyen and Ngoc Hoang Luong employed a widely-used multi-objective evolutionary algorithm, i.e., the non-dominated sorting genetic algorithm II (NSGA-II), to approximate the optimal Pareto-optimal fronts for Neural Architecture Search problem formulations. Experimental results on the NAS benchmark NATS-Bench show the advantages and disadvantages of each formulation.

The second paper, "An Automatic Labeling Method for Subword-Phrase Recognition in Effective Text Classification" by Yusuke Kimura, Takahiro Komamizu, and Kenji Hatano, proposed new technique to add subword-phrase recognition as an auxiliary task and utilizing it for text classification.

The third paper, "Motion Embedded Images: An Approach to Capture Spatial and Temporal Features for Action Recognition", by Tri Le, Nham Huynh-Duc, Chung Thai Nguyen and Minh-Triet Tran, investigated the use of motion-embedded images in a variant of two-stream Convolutional Neural Network architecture, in which one stream captures motion using combined batches of frames, while another stream employs a normal image classification ConvNet to classify static appearance.

The fourth paper, "Complaints with Target Scope Identification on Social Media", by Kazuhiro Ito, Taichi Murayama, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki, benchmark the annotated Japanese text dataset by machine learning baselines and obtain the best performance of 90.4 F1-score in detecting whether a text was a complaint or not, and a micro-F1 score of 72.2 in identifying the target scope label. This paper experimented on these models to demonstrate that identifying a target scope of complaints is useful for sociological analysis.

The fifth paper, "Khmer-Vietnamese Neural Machine Translation Improvement Using Data Augmentation Strategies", by Thai Nguyen Quoc and Huong Le Thanh, employs a pre-trained multilingual model and fine-tunes it by using a small bilingual dataset. The proposed approach is applied to the Khmer-Vietnamese machine translation.

The last paper, "A Hybrid Deep Learning Approach to Keyword Spotting in Vietnamese Stele Images", by Anna Scius-Bertrand, Marc Bui, and Andreas Fischer, presents a hybrid approach to spot keywords in stele images that combines data-driven deep learning with knowledge-based structural modeling and matching of Chu Nom characters.

Guest Editors

Huynh Thi Thanh Binh
(binhht@soict.hust.edu.vn), School of Information and Communication Technology
Hanoi University of Science and Technology, Japan

Ichiro Ide
(ide@i.nagoya-u.ac.jp), Nagoya University, Graduate School of Informatics / Mathematical & Data Science Center, Japan

Lightweight Multi-Objective and Many-Objective Problem Formulations for Evolutionary Neural Architecture Search with the Training-Free Performance Metric Synaptic Flow

An Vo, Tan Ngoc Pham, Van Bich Nguyen and Ngoc Hoang Luong
University of Information Technology, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
E-mail: 19520007@gm.uit.edu.vn, 19520925@gm.uit.edu.vn, vannb@uit.edu.vn, hoangln@uit.edu.vn

Keywords: neural architecture search, evolutionary algorithms, multi-objective optimization

Received: March 11, 2023

Neural architecture search (NAS) with naïve problem formulations and applications of conventional search algorithms often incur prohibitive search costs due to the evaluations of many candidate architectures. For each architecture, its accuracy performance can be properly evaluated after hundreds (or thousands) of computationally expensive training epochs are performed to achieve proper network weights. A so-called zero-cost metric, Synaptic Flow, computed based on random network weight values at initialization, is found to exhibit certain correlations with the neural network test accuracy and can thus be used as an efficient proxy performance metric during the search. Besides, NAS in practice often involves not only optimizing for network accuracy performance but also optimizing for network complexity, such as model size, number of floating point operations, or latency, as well. In this article, we study various NAS problem formulations in which multiple aspects of deep neural networks are treated as multiple optimization objectives. We employ a widely-used multi-objective evolutionary algorithm, i.e., the non-dominated sorting genetic algorithm II (NSGA-II), to approximate the optimal Pareto-optimal fronts for these NAS problem formulations. Experimental results on the NAS benchmark NATS-Bench show the advantages and disadvantages of each formulation.

Povzetek: Uporabljen je algoritem NSGA-II za analizo NAS problemov, tj. za iskanje primerne nevronske arhitekture.

1 Introduction

The goal of Neural Architecture Search (NAS) is to accelerate the design process of high-performing deep neural network architectures by exploring the vast space of possible network configurations and selecting the most promising ones. This process typically involves searching over a large number of potential architectures, evaluating their performance, and iteratively refining the algorithm to converge on the best-performing architectures [12]. However, many state-of-the-art NAS methods require substantial computational resources. For example, Zoph et al. [30] employed 800 GPUs over 28 days to solve NAS using a reinforcement earning algorithm, whereas Real et al. [27] proposed an evolution-based technique (AmoebaNet-A) that took 7 days to execute on 450 K40 GPUs. To reduce such heavy computation costs, current NAS efficiency research proposes the adoption of *training-free performance metrics* [1] as a performance objective rather than network accuracy. These metrics can be computed using network weights at initialization and do not require any training epochs, thus primarily involving network designs. Several such training-free metrics have been shown to be correlated with actual network accuracy to some extent [1]. Hence, optimizing these metrics potentially leads to promising architectures.

While most studies focus on optimizing network architectures for a single objective, such as network accuracy, real-world neural network deployments frequently necessitate the consideration of other important factors, such as FLOPs, number of parameters, and latency. NAS architectures that are optimized just for accuracy may be too cumbersome for resource-constrained embedded devices. Moreover, by solving multi-objective problems, a trade-off front between performance and complexity can be obtained, which provides decision-makers with the necessary information to select an appropriate network. Several research has presented multi-objective NAS (MONAS) formulations that take into consideration important aspects. For example, Lu et al. [20] presented NSGA-Net, which used the non-dominated sorting genetic algorithm II (NSGA-II) [6] to solve an MONAS problem with two conflicting objectives, i.e., the classification error and the number of floating-point operations (FLOPs). In another work [19], NSGA-II was also used to solve a many-objective problem formulation with five optimization objectives, including ImageNet accuracy, number of parameters, number of multiply-add operations, CPU and GPU latency.

Lu et al. [19] also developed a surrogate model to forecast the accuracy of candidate architectures and the predictor was refined during the search process to enhance the performance of NSGA-II in solving MONAS. To build the predictor, a limited number of architectures were sampled from the search space at first. Following that, NSGA-II was used to search for architectures, treating the accuracy predictor as an objective alongside other complexity objectives. Despite the fact that they employed a surrogate model as an objective for NSGA-II to discover architectures, they still trained these architectures and used them as training samples to refine the accuracy predictor. Using complexity metrics and training-free performance metric Synaptic Flow (`synflow`) simultaneously, Phan et al. [25] randomly choose a wide variety of architectures and evaluate their complexity and performance. Non-dominated architectures with high performance and low complexity are then utilized to initialize the population for a bi-objective evolutionary NAS process where network accuracy is used as the primary performance metric. The training-free `synflow` metric is only employed during the *warm-up* phase. During the search phase, candidate architectures still need to be trained and evaluated for their performance. It's also possible to use `synflow` metric to enhance the performance of NSGA-II in solving multi-objective NAS problems as in [26], by developing a training-free multi-objective local search. In each generation, a subset of potential architectures undergoes a local search process that uses `synflow` for improvement checks, eliminating the need for training epochs. In contrast to these works, our approach does not rely on any training process. Instead, we use the training-free performance metric `synflow` to evaluate all candidate architectures during the search. This eliminates the need for training and allows us to search for high-quality architectures more efficiently. Do et al. [7] also propose a completely training-free multi-objective evolutionary NAS framework that employs the number of linear regions $R_{\mathcal{N}}$ and the condition number of the neural tangent kernel $\kappa_{\mathcal{N}}$ to evaluate candidate architectures, which are data-dependent metrics computed using mini-batches from a training dataset. In our work, we use the data-agnostic metric `synflow` as our performance objective. The resulting architectures are thus potentially applicable to a wider range of tasks and datasets.

This article extends our SoICT 2022 conference paper on training-free multi-objective and many-objective evolutionary NAS [29]. In [29], we discussed several multi-objective and many-objective NAS problem formulations and employed the well-known multi-objective evolutionary algorithm NSGA-II to solve these formulations. Moreover, we exclusively used the data-agnostic training-free metric `synflow` to evaluate candidate architecture performance without any training. In this article, we extend the analysis in our preliminary work by adding the hypervolume performance indicator results instead of only Inverted Generational Distance (IGD). While IGD exhibits the convergence behavior of a multi-objective algorithm, it cannot be used in real-world situations due to its requirement of the Pareto-optimal front (see Sections 2.1 and 4.2.1). The hypervolume, which requires only a reference nadir point, is a more practical performance indicator for evaluating and comparing multi-objective NAS approaches (see Section 4.2.2). Employing both IGD and hypervolume thus yields more detailed investigations into the effectiveness of different NAS problem formulations. We present the IGD and hypervolume results in terms of GPU hours rather than the number of generations, which allows us to better assess the efficiency of our approaches. Our experimental results demonstrate that Training-Free Many-Objective Evolutionary NAS (TF-MaOENAS) provides several advantages when achieving competitive results while taking only 3 GPU hours.

2 Backgrounds

2.1 Multi-objective neural architecture search

Multi-Objective NAS (MONAS) [20, 26] can be formulated as searching for high-quality architectures in a search space Ω where m different aspects (e.g., error rate, model size, or latency) are optimized simultaneously. Each aspect is modeled as a separate objective $f_i(\boldsymbol{x})$, $i \in \{1, \ldots, m\}$, and each candidate architecture $\boldsymbol{x} \in \Omega$ thus has a corresponding vector of objective values $\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x}))$. All objectives, without loss of generality, are assumed to be minimized.

An architecture $\boldsymbol{x}$ *dominates* another architecture $\boldsymbol{y}$ if and only if $\boldsymbol{x}$ strictly outperforms $\boldsymbol{y}$ in at least one aspect and $\boldsymbol{x}$ is never outperformed by $\boldsymbol{y}$ in any aspects:

$$\boldsymbol{x} \prec \boldsymbol{y} \iff \forall i, f_i(\boldsymbol{x}) \leq f_i(\boldsymbol{y}) \land \exists i, f_i(\boldsymbol{x}) < f_i(\boldsymbol{y})$$

If some objectives conflict with each other, e.g., network prediction accuracy versus network complexity, there will not exist an *ideal* architecture optimizing all those competing objectives. Instead, there exists a ***Pareto set*** P_S of architectures, in which all can be considered Pareto-optimal because they are not dominated by any other architectures:

$$P_S = \{\boldsymbol{x} \in \Omega \mid \nexists \boldsymbol{x}' \in \Omega, \boldsymbol{x}' \prec \boldsymbol{x}\}$$

The images of all Pareto-optimal architectures in P_S together form a ***Pareto-optimal front*** P_F in the objective space [5, 23]:

$$P_F = \{\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^m \mid \boldsymbol{x} \in P_S\}$$

Each point on P_F denotes the vector of objective values $\boldsymbol{f}(\boldsymbol{x})$ of a Pareto-optimal architecture $\boldsymbol{x}$, which exhibits an optimal trade-off among the competing objectives. For example, from a Pareto-optimal architecture $\boldsymbol{z}$, if we modify $\boldsymbol{z}$ to improve network performance (e.g., prediction accuracy), network complexity (e.g., model size or FLOPs) must be increased as well. In other words, there exists no means in the search space Ω to alter $\boldsymbol{z}$ in order to increase accuracy performance without incurring additional computation cost. A Pareto-optimal front P_F thus exhibits insightful information for decision-makers, e.g., which Pareto-optimal architecture on P_F exhibits the most desirable trade-off between network latency and accuracy.

The optimal solution of MONAS is not a single ideal architecture but the Pareto set P_S. However, achieving the entire P_S is prohibitively costly (if there are many Pareto-optimal architectures) and unnecessary (if choosing between architectures close to each other on the Pareto-optimal front P_F does not make considerable differences). It is often more practical to find an *approximation set* S that yields a good *approximation front* $\boldsymbol{f}(S)$ well approximating the Pareto-optimal front P_F in terms of both proximity and diversity [6, 21].

2.2 Non-dominated sorting genetic algorithm II

Evolutionary Algorithms (EAs) are often employed for handling multi-objective optimization problems because their intrinsic population-based operations are well-suited for the goal of finding multiple non-dominated solutions to approximate Pareto-optimal fronts [5, 15, 23]. In this article, we consider the Non-dominated Sorting Genetic Algorithm II (NSGA-II) [6] as the optimization algorithm. NSGA-II has also been widely used for solving different MONAS problem instances as well [19, 20]. In the following paragraph, we provide a brief description of NSGA-II, and further details can be found in [6].

The NSGA-II population P is initialized with N individuals, where each individual corresponds with a candidate architecture randomly sampled from the search space Ω. Until the computation budget is over, or other termination criteria are met, NSGA-II operates in a generational manner as follows. In every generation, a set S of N promising individuals in terms of Pareto dominance are selected from P via binary tournament selection. N new candidate architectures (i.e., set O of offspring individuals) are generated from the parent architectures (i.e., set S of selected individuals) via variation operators (i.e., crossover and mutation) and are evaluated for their objective values. The current population P and the offspring population are then merged into a pool $(P+O)$ where all $2N$ individuals are sorted into their *non-domination ranks* $0, 1, 2, \ldots$. Rank 0 consists of individuals that are not dominated by any other individuals in $(P + O)$, and rank i consists of individuals that would be non-dominated if individuals from lower ranks ($< i$) are omitted. A group of N promising individuals are then selected from the pool $(P + O)$ via truncation selection to form the population for the next generation. Lower-rank individuals are selected first, and if selections need to be performed among individuals of the same rank, far-apart individuals are preferred.

2.3 Training-free performance metric synaptic flow

Synaptic Flow (`synflow`) is a metric for measuring the importance of each parameter in a neural network architecture, based on the inter-layer interaction of other network parameters. Tanaka et al. [28] first introduced the `synflow` score for single parameter $w_{ij}^{[l]}$ in the l-th layer of a fully-connected neural network as follows:

$$\mathcal{P}(w_{ij}^{[l]}) = \left[\mathbf{1}^T \prod_{k=l+1}^{N} \left|W^{[k]}\right|\right]_i \left|w_{ij}^{[l]}\right| \left[\prod_{k=1}^{l-1} \left|W^{[k]}\right| \mathbf{1}\right]_j \quad (1)$$

where $W^{[k]}$ is the weight matrix of the k-th layer of the network. The `synflow` score for a parameter $w_{ij}^{[l]}$ takes into account the product of the absolute values of the weights of all the layers downstream from the current layer l, as well as the product of the absolute values of the weights of all the layers upstream from the current layer l. Thus, the `synflow` score of a parameter $\mathcal{P}_S(w_{ij}^{[l]})$ reflects the contribution of that parameter to the information flow of the network.

Abdelfattah et al. [1] then extended the `synflow` score to evaluate the entire network architecture $\boldsymbol{x}$, which is sum of the `synflow` scores for all M parameters in the network as follows:

$$\mathcal{S}(\boldsymbol{w}(\boldsymbol{x})) = \sum_{i=1}^{M} \mathcal{P}(w_i(\boldsymbol{x})) \quad (2)$$

According to [1], the training-free performance metric `synflow` exhibits a strong correlation with the final accuracy of the network in the NAS-Bench-201 architecture search space, with Spearman ρ coefficients of 0.74, 0.76, and 0.75 on CIFAR-10, CIFAR-100, and ImageNet16-120, respectively. The bi-objective space of test accuracy after 200 training epochs versus FLOPs for all architectures in NATS-Bench is depicted in Figure 1. According to the graph, architectures with greater `synflow` scores tend to have higher test accuracy. Furthermore, `synflow` is a data-agnostic metric that can be computed solely based on network weights (see Equation 1). Unlike other performance metrics that do not require training, such as `jacob_cov` [22] or the condition number of the NTK [4], `synflow` does not need any data mini-batches to be used as input for the network. It means that `synflow` can be used to measure the performance of a neural network without having to pass any data through it. This is beneficial since it allows for a more efficient evaluation of a neural network's performance without having to expend resources on data collection and pre-processing. Since `synflow` is data-independent and does not quire any training epochs, it can serve as an effective proxy for optimizing network accuracy in tackling NAS problems [1].

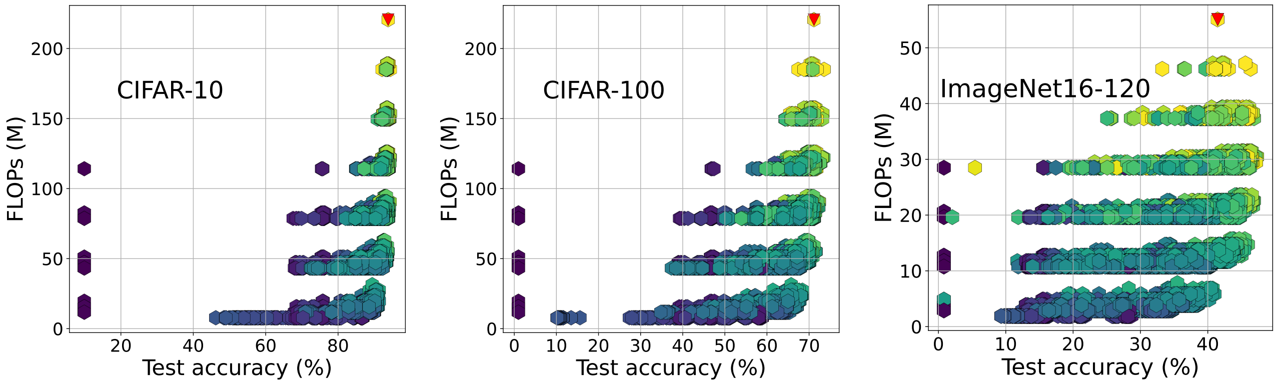


Figure 1: Illustration of all network architectures in the NATS-Bench search space. Brighter hexagons indicate greater values of `synflow`, while red triangles denote the architectures with the highest `synflow` values.

3 NAS problem formulations

3.1 Multi-objective NAS problem formulation

NAS can be formulated as a multi-objective optimization problem, which seeks to simultaneously optimize two objectives, such as accuracy and computational complexity, as follows:

$$\begin{aligned}\min \quad & \boldsymbol{F}(\boldsymbol{x}) = (f_{\text{err}}(\boldsymbol{x}, \boldsymbol{w}^*(\boldsymbol{x}), \mathcal{D}_{\text{val}}), f_{\text{comp}}(\boldsymbol{x})) \in \mathbb{R}^2, \\ \text{st} \quad & \boldsymbol{w}^*(\boldsymbol{x}) \in \arg\min \mathcal{L}(\boldsymbol{x}, \boldsymbol{w}(\boldsymbol{x}), \mathcal{D}_{\text{train}}), \\ & \boldsymbol{x} \in \Omega_{\text{arch}}, \quad \boldsymbol{w}(\boldsymbol{x}) \in \Omega_{\text{weight}}(\boldsymbol{x}),\end{aligned} \tag{3}$$

where $\boldsymbol{x}$ denotes an architecture in the search space Ω_{arch}. Multi-Objective NAS (MONAS), aiming to find a set of architectures that exhibit optimal trade-offs between accuracy and complexity, is a bi-level optimization problem. At the upper level, it seeks high-quality candidate architectures that optimize both error rate f_{err} and complexity f_{comp}, while at the lower level, it searches for the proper network weight values $\boldsymbol{w}^*(\boldsymbol{x})$ for each candidate architecture $\boldsymbol{x}$. The network weight values $\boldsymbol{w}(\boldsymbol{x})$ must be specified in order to accurately evaluate the error rate of a network architecture $\boldsymbol{x}$. This requires solving a lower-level optimization problem over the network weight space $\Omega_{\text{weight}}(\boldsymbol{x})$ of the given architecture $\boldsymbol{x}$. By doing so, we can obtain a set of weight values that minimize the error rate and maximize the performance of the network. This is typically done by employing a stochastic gradient descent (SGD) algorithm to perform many iterative updates on network weight values in order to minimize a loss function $\mathcal{L}$, which measures the difference between network predictions and ground-truth targets for data items in a training dataset D_{train}. This process can be computationally expensive and time-consuming, but is necessary in order to accurately obtain the proper values of $\boldsymbol{w}(\boldsymbol{x})$ for any given architecture.

The two optimization objectives at the upper level of MONAS are minimizing error rate f_{err} and minimizing complexity f_{comp}. The complexity of a network can be assessed via metrics such as the number of floating point operations (FLOPs), latency, the number of parameters (#parameters), or the number of multiply-accumulate units (#MACs). These metrics can be calculated without the weights of the network, and thus require minimal computing time. Besides, in order to prevent overfitting to the training dataset D_{train}, it is important to calculate error rate f_{err} on a separate validation dataset D_{val} that has not been used for training. At the end of a search, the error rates and weight of resulting architectures should be tested on a new dataset D_{test} to measure their ability to generalize.

3.2 Training-free multi-objective NAS problem formulation

Evaluating the prediction performance of multiple candidate architectures in MONAS requires computationally intensive training procedures (i.e., lower-level optimization) which consume a significant amount of computing time (see Equation 3). To eliminate this cost, several NAS formulations have been proposed to use training-free performance metrics as proxies for the network error rate f_{err}. This approach allows us to quickly evaluate candidate architectures without having to perform costly training procedures. We present a training-free bi-objective NAS formulation that uses the `synflow` metric as an alternative to f_{err} as follows:

$$\begin{aligned}\text{minimize} \quad & \boldsymbol{F}(\boldsymbol{x}) = (f_{\text{SF}}(\boldsymbol{x}, \boldsymbol{w}(\boldsymbol{x})), f_{\text{comp}}(\boldsymbol{x})) \in \mathbb{R}^2, \\ \text{subject to} \quad & \boldsymbol{x} \in \Omega_{\text{arch}}, \quad \boldsymbol{w}(\boldsymbol{x}) \in \Omega_{\text{weight}}(\boldsymbol{x}),\end{aligned} \tag{4}$$

where $f_{\text{SF}}(\boldsymbol{x}, \boldsymbol{w}(\boldsymbol{x})) = -\mathcal{S}(\boldsymbol{w}(\boldsymbol{x}))$ (as `synflow` should be maximized). At the start of the lower level optimization process, $\boldsymbol{w}(\boldsymbol{x})$ can be initialized randomly to compute their `synflow` scores. We name this formulation TF-MONAS.

3.3 Training-free many-objective NAS problem formulation

The TF-MONAS formulation in Equation 4 can be further extended by incorporating many objectives simultaneously. This approach allows for the simultaneous consideration of multiple complexity metrics, such as FLOPs, latency, #parameters, and #MACs. By optimizing the neural network architecture from various perspectives, this approach can ensure that the resultant architecture is suitable for a variety of applications. A common scenario is the use of deep neural networks on embedded devices, such as drones or smart watches. This requires taking into account a variety of hardware limitations, such as model size and storage capacity, as well as usage requirements like response time. All of these must be considered when deploying deep neural networks on embedded devices in order to ensure reliable performance. Additionally, we note that these complexity metrics can be evaluated without incurring too much computing cost. We formulate training-free many-objective evolutionary NAS (TF-MaOENAS) that does not require any training and consists of five objectives as follows:

$$\begin{aligned} \text{minimize} \quad & \boldsymbol{F}(\boldsymbol{x}) = (f_{\text{SF}}(\boldsymbol{x}, \boldsymbol{w}(\boldsymbol{x})), f_{\text{MACs}}(\boldsymbol{x}), \\ & f_{\text{latency}}(\boldsymbol{x}), f_{\text{FLOPs}}(\boldsymbol{x}), f_{\text{params}}(\boldsymbol{x})), \\ \text{subject to} \quad & \boldsymbol{x} \in \Omega_{\text{arch}}, \quad \boldsymbol{w}(\boldsymbol{x}) \in \Omega_{\text{weight}}(\boldsymbol{x}), \end{aligned} \tag{5}$$

The complexity metrics used in this work are FLOPs , latency, the number of parameters (#parameters), and #MACs. We name this formulation TF-MaOENAS. We will compare solving one TF-MaOENAS formulation that involves four optimization objectives for network complexity against solving separately four different TF-MOENAS formulations in which each model considers only one complexity objective. Moreover, we also compare TF-MaOENAS with (training-based) MaOENAS to demonstrate the benefits of the training-free performance metric `synflow`.

4 Experiments

4.1 Experimental details

Our experiments are carried out on NATS-Bench [8], which is an extended version of NAS-Bench-201 [11]. NATS-Bench comprises 15,625 architectures and provides a variety of metrics for evaluating different architectures, such as accuracy, number of parameters, and training time, across three datasets: CIFAR-10, CIFAR-100, and ImageNet16-120. We experiment with four MONAS approaches on NATS-Bench, each with specific optimization objectives as follows:

1. **04 MOENAS variants**: 01 training-based performance metric (validation accuracy after 12 training epochs as in other related works [11, 26]) versus 01 complexity metric (FLOPs, #parameters, latency, or #MACs).
2. **04 TF-MOENAS variants**: 01 training-free performance metric (`synflow`) versus 01 complexity metric (FLOPs, #parameters, latency, or #MACs).
3. **01 MaOENAS variant**: 01 training-based performance metric (validation accuracy after 12 training epochs) versus 04 complexity metrics (FLOPs, latency, #parameters, and #MACs).
4. **01 TF-MaOENAS variant**: 01 training-free performance metric (`synflow`) versus 04 complexity metrics (FLOPs, latency, #parameters, and #MACs).

We use the NSGA-II [6] as our multi-objective search algorithm. We set the population size to 20, the number of generations to 50, and used random initialization. We also employ binary tournament selection, two-point crossover with a probability of 0.9, and polynomial mutation with a probability of $1/l$, where l is the encoding length of each individual.

Besides, we also implement an *elitist archive* [21] to save non-dominated architectures discovered so far throughout the NAS process. When an architecture is evaluated, it is included in the elitist archive if it is not dominated by any existing architectures in the elitist archive. Existing architectures that are dominated by newly added architecture will be removed from the elitist archive. The non-dominated architectures in the elitist archive, therefore, constitute an approximation set, which may be regarded as the NSGA-II optimization result. The elitist archive is only used for result logging and does not impact the workings of NSGA-II. Because non-dominated solutions might be lost due to the stochasticity of the variation and selection operators, an elitist archive is desirable for multi-objective evolutionary algorithms.

We conduct 21 independent runs of NSGA-II for each problem formulation presented in Section 3 on CIFAR-10, CIFAR-100, and ImageNet16-120 of NATS-Bench. All of our experiments can be performed using Google Colab.

4.2 Performance metric

4.2.1 Inverted generational distance

To compare an approximation set S of non-dominated architectures against the Pareto-optimal front P_F of the most efficient trade-off architectures, we employ the *Inverted Generational Distance* (IGD) [3] which is defined as:

$$\text{IGD}(\mathcal{S}, P_F) = \frac{1}{|P_F|} \sum_{p \in P_F} \min_{\boldsymbol{x} \in \mathcal{S}} \|p - \boldsymbol{f}(\boldsymbol{x})\|_2 \tag{6}$$

The smaller IGD indicates the better approximation front achieved by the current solutions. For example, if $\text{IGD}(\mathcal{S}_1, P_F) < \text{IGD}(\mathcal{S}_2, P_F)$, then $\mathcal{S}_1$ is a better approximation front compared to $\mathcal{S}_2$ regarding P_F. The Pareto-optimal front P_F can be obtained by iterating over all architectures in the NAS benchmark. The Pareto-optimal front P_F is computed by querying the database of NATS-Bench for test accuracy values after 200 epochs. Approximation sets $\mathcal{S}$ are taken from the elitist archive obtained from the search process. The IGD value between the archive and the Pareto-optimal front P_F can be calculated after each evolutionary generation to measure how close the current approximation front of the algorithm is to the front of

Pareto-optimal architectures P_F. The test accuracy after 200 epochs and IGD values are only used to assess the effectiveness of the search algorithms and are not employed to direct the search process.

4.2.2 Hypervolume

Hypervolume [3, 18] is also a measure of the quality of a set of non-dominated solutions in multi-objective optimization besides IGD. It can be computed by measuring the area covered by the solution points on the approximation front with regard to a *reference point*. In contrast to IGD, which requires the Pareto-optimal front P_F to compute (IGD can thus hardly be used in real-world multi-objective optimization), hypervolume only need a reference point to be specified, which is usually the nadir point (the worst point in the objective space). The higher hypervolume implies that the corresponding method achieves a better approximation front. For example, if Hypervolume$(\mathcal{S}_1) >$ Hypervolume$(\mathcal{S}_2)$, then $\mathcal{S}_1$ is a better approximation front compared to $\mathcal{S}_2$.

4.3 Result analysis

	IGD	Hypervolume	Test accuracy	Search cost (hours)
Space: Test accuracy - FLOPs				
(1)	0.0198 ± 0.0171	1.0332 ± 0.0013	94.28 ± 0.17	53.7
(2)	0.0250 ± 0.0133	1.0223 ± 0.0027	94.29 ± 0.17	0.7
(3)	0.0308 ± 0.0177	$\underline{1.0334 \pm 0.0011}$	94.27 ± 0.13	54.8
(4)	$\underline{0.0096 \pm 0.0021}$	1.0298 ± 0.0022	$\underline{94.37 \pm 0.00}$	2.7
Space: Test accuracy - Latency				
(1)	$\underline{\mathbf{0.0228 \pm 0.0019}}$	$\underline{\mathbf{1.0050 \pm 0.0006}}$	94.30 ± 0.09	54.1
(2)	0.0532 ± 0.0056	0.9431 ± 0.0200	94.29 ± 0.14	1.1
(3)	0.0277 ± 0.0060	0.9967 ± 0.0168	94.27 ± 0.13	54.8
(4)	0.0412 ± 0.0060	0.9581 ± 0.0098	$\underline{94.37 \pm 0.00}$	2.7
Space: Test accuracy - #Parameters				
(1)	0.0180 ± 0.0138	1.0332 ± 0.0014	94.27 ± 0.18	53.8
(2)	0.0314 ± 0.0170	1.0233 ± 0.0027	94.24 ± 0.22	0.8
(3)	0.0309 ± 0.0176	$\underline{1.0334 \pm 0.0011}$	94.27 ± 0.13	54.8
(4)	$\underline{0.0098 \pm 0.0022}$	1.0296 ± 0.0023	$\underline{94.37 \pm 0.00}$	2.7
Space: Test accuracy - #MACs				
(1)	0.0195 ± 0.0131	1.0331 ± 0.0017	94.24 ± 0.22	53.8
(2)	0.0189 ± 0.0069	1.0280 ± 0.0034	94.35 ± 0.03	0.8
(3)	0.0266 ± 0.0150	$\underline{1.0333 \pm 0.0011}$	94.27 ± 0.13	54.8
(4)	$\underline{\mathbf{0.0104 \pm 0.0023}}$	1.0292 ± 0.0025	$\underline{94.37 \pm 0.00}$	2.7

Table 1: Results of search and evaluation directly on CIFAR-10: (1) MOENAS, (2) TF-MOENAS, (3) MaOENAS, (4) TF-MaOENAS. Results that are underlined indicate the best method and results that are **bolded** denote the best method with statistical significance (p-value < 0.01)

	IGD	Hypervolume	Test accuracy	Search cost (hours)
Space: Test accuracy - FLOPs				
(1)	0.0384 ± 0.0086	0.7958 ± 0.0015	72.39 ± 0.21	53.8
(2)	0.0493 ± 0.0176	0.7851 ± 0.0036	72.56 ± 0.44	0.8
(3)	0.0334 ± 0.0128	0.7964 ± 0.0019	72.40 ± 0.30	54.8
(4)	$\underline{\mathbf{0.0122 \pm 0.0045}}$	$\underline{\mathbf{0.7993 \pm 0.0019}}$	$\underline{73.49 \pm 0.07}$	2.7
Space: Test accuracy - Latency				
(1)	$\underline{0.0318 \pm 0.0070}$	$\underline{0.7960 \pm 0.0013}$	72.68 ± 0.68	54.0
(2)	0.1182 ± 0.0139	0.7460 ± 0.0149	$\underline{73.51 \pm 0.00}$	1.0
(3)	0.0352 ± 0.0084	0.7701 ± 0.0057	72.40 ± 0.30	54.8
(4)	0.0446 ± 0.0057	0.7539 ± 0.0076	73.49 ± 0.07	2.7
Space: Test accuracy - #Parameters				
(1)	0.0369 ± 0.0029	0.7960 ± 0.0013	72.47 ± 0.23	53.8
(2)	0.0189 ± 0.0038	0.7883 ± 0.0025	73.47 ± 0.11	0.8
(3)	0.0335 ± 0.0127	0.7963 ± 0.0019	72.40 ± 0.30	54.8
(4)	$\underline{\mathbf{0.0123 \pm 0.0045}}$	$\underline{\mathbf{0.7990 \pm 0.0020}}$	$\underline{73.49 \pm 0.07}$	2.7
Space: Test accuracy - #MACs				
(1)	0.0313 ± 0.0094	0.7956 ± 0.0021	72.39 ± 0.36	53.8
(2)	0.0156 ± 0.0025	0.7941 ± 0.0041	$\underline{73.51 \pm 0.00}$	0.8
(3)	0.0270 ± 0.0096	0.7961 ± 0.0018	72.40 ± 0.30	54.8
(4)	$\underline{\mathbf{0.0126 \pm 0.0041}}$	$\underline{0.7985 \pm 0.0021}$	73.49 ± 0.07	2.7

Table 2: Results of search and evaluation directly on CIFAR-100: (1) MOENAS, (2) TF-MOENAS, (3) MaOENAS, (4) TF-MaOENAS. Results that are underlined indicate the best method and results that are **bolded** denote the best method with statistical significance (p-value < 0.01)

Figure 2 demonstrates that TF-MaOENAS achieves superior IGD convergence results compared to other approaches while taking just 3 GPU hours in most cases, with the exception of test accuracy versus latency space. However, in terms of hypervolume, MaOENAS and MOENAS alternatively surpass other approaches on CIFAR-10 and ImageNet16-120. Table 1, Table 2, and Table 3 show comprehensive results on CIFAR-10, CIFAR-100 and ImageNet16-120. It is noted that the hypervolume of TF-MaOENAS still outperforms other methods in the majority of cases on CIFAR-100, and its hypervolume is only slightly lower than that of other training-based methods on CIFAR-10 and ImageNet16-120. Furthermore, because it is a training-free approach, it only requires 3 GPU hours as opposed to dozens to hundreds of GPU hours for training-based methods like MOENAS and MaOENAS. Regarding test accuracy, TF-MaOENAS discovers top-performing architectures on NATS-Bench and outperforms other methods in the majority of situations.

The experimental results also show that TF-MaOENAS and TF-MOENAS (using `synflow`) perform better than MaOENAS and MOENAS (using validation accuracy after 12 training epochs), respectively. This indicates that using `synflow` is more effective at optimizing for multiple objectives simultaneously than using the validation accuracy after 12 training epochs. This might reflect that the training-free `synflow` metric is more capable of measuring and balancing between optimizing for accuracy and other complexity objectives. Moreover, `synflow` is a training-free metric, it just takes a few seconds to compute, resulting in a lower computing cost than a training-based metric. On the other hand, TF-MaOENAS, which employs five objectives concurrently, outperforms TF-MOENAS, which employs only two objectives. This is due to the addition of MACs, the number of parameters, and latency as complexity objectives in addition to `synflow` and FLOPs. Most of the time, optimizing more objectives is favorable while not incurring considerably more computing expenses. This will provide a fuller picture of the complexity of achieved architectures, enabling a more precise evaluation of the trade-offs between performance and complexity. Additionally, the penta-objective approximation fronts obtained by TF-MaOENAS can be projected into different bi-objective spaces (i.e., test accuracy versus one complexity metric) and still achieve better results than the corresponding TF-MOENAS variants. This means that running TF-MaOENAS once in the penta-objective space can obtain good approximation fronts in different bi-objective

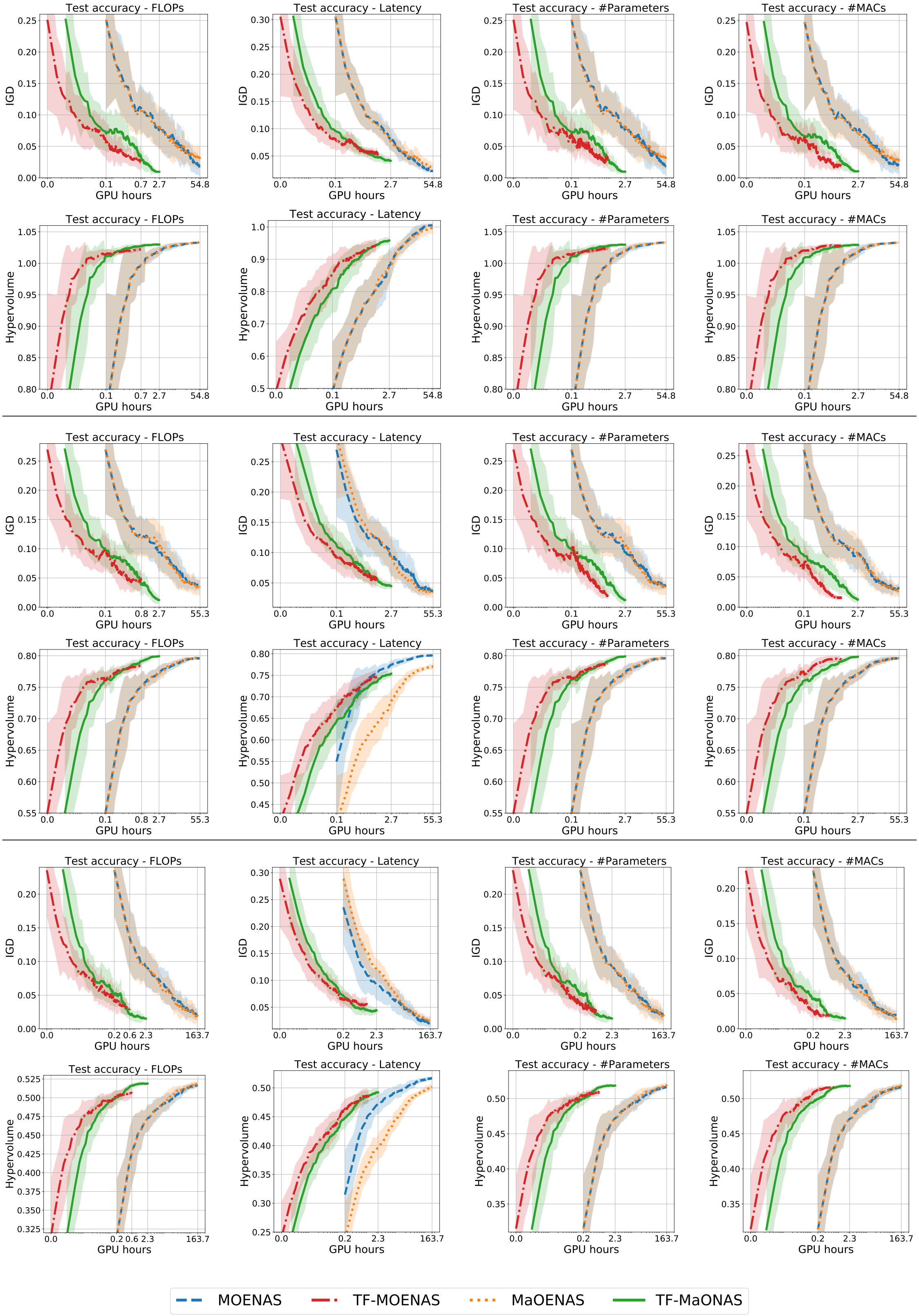


Figure 2: IGD and hypervolume comparisons in terms of GPU hours (log scale) on four different bi-objective spaces (plot title) across CIFAR-10 (top two rows), CIFAR-100 (middle two rows) and ImageNet16-120 (bottom two rows). The figures depict the mean values with lines and the standard deviation with shaded areas over 21 runs.

	IGD	Hypervolume	Test accuracy	Search cost (hours)
	Space: Test accuracy - FLOPs			
(1)	0.0217 ± 0.0087	0.5165 ± 0.0026	46.34 ± 0.35	161.8
(2)	0.0296 ± 0.0089	0.5062 ± 0.0062	46.25 ± 0.15	0.6
(3)	0.0192 ± 0.0165	0.5193 ± 0.0032	46.41 ± 0.43	163.7
(4)	0.0151 ± 0.0019	0.5189 ± 0.0017	46.57 ± 0.05	2.3
	Space: Test accuracy - Latency			
(1)	0.0281 ± 0.0047	0.5171 ± 0.0019	46.62 ± 0.52	162.2
(2)	0.0543 ± 0.0112	0.4852 ± 0.0118	46.52 ± 0.17	1.1
(3)	0.0192 ± 0.0165	0.5012 ± 0.0059	46.41 ± 0.43	163.7
(4)	0.04428 ± 0.0062	0.4922 ± 0.0086	46.57 ± 0.05	2.3
	Space: Test accuracy - #Parameters			
(1)	0.0194 ± 0.0067	0.5171 ± 0.0019	46.46 ± 0.24	161.8
(2)	0.0264 ± 0.0114	0.5092 ± 0.0047	46.40 ± 0.18	0.8
(3)	0.0194 ± 0.0165	0.5191 ± 0.0032	46.41 ± 0.43	163.7
(4)	0.0153 ± 0.0019	0.5186 ± 0.0017	46.57 ± 0.05	2.3
	Space: Test accuracy - #MACs			
(1)	0.0198 ± 0.0073	0.5153 ± 0.0039	46.17 ± 0.46	161.8
(2)	0.0188 ± 0.0020	0.5161 ± 0.0020	46.57 ± 0.02	0.6
(3)	0.0156 ± 0.0107	0.5188 ± 0.0032	46.41 ± 0.43	163.7
(4)	**0.0148 ± 0.0022**	0.5181 ± 0.0017	46.57 ± 0.05	2.3

Table 3: Results of search and evaluation directly on ImageNet16-120: (1) MOENAS, (2) TF-MOENAS, (3) MaOENAS, (4) TF-MaOENAS. Results that are underlined indicate the best method and results that are **bolded** denote the best method with statistical significance (p-value < 0.01)

Alg.	CIFAR-10 (direct)	CIFAR-100 (transfer)	ImageNet16-120 (transfer)	Search cost (hours)
	Space: Test accuracy - FLOPs			
(1)	0.0198 ± 0.0171	0.0465 ± 0.0183	0.0316 ± 0.0147	53.7
(2)	0.0250 ± 0.0133	0.0322 ± 0.0103	0.0400 ± 0.0145	0.7
(3)	0.0308 ± 0.0177	0.0299 ± 0.0106	0.0230 ± 0.0091	54.8
(4)	0.0096 ± 0.0021	**0.0125 ± 0.0017**	**0.0161 ± 0.0016**	2.7
	Space: Test accuracy - Latency			
(1)	**0.0228 ± 0.0019**	0.0419 ± 0.0056	0.0416 ± 0.0103	54.1
(2)	0.0532 ± 0.0056	0.0932 ± 0.0100	0.0841 ± 0.0175	1.1
(3)	0.0277 ± 0.0060	0.0390 ± 0.0093	0.0369 ± 0.0049	54.8
(4)	0.0412 ± 0.0060	0.0612 ± 0.0091	0.0577 ± 0.0097	2.7
	Space: Test accuracy - #Parameters			
(1)	0.0180 ± 0.0138	0.0413 ± 0.0171	0.0342 ± 0.0139	53.8
(2)	0.0314 ± 0.0170	0.0502 ± 0.0144	0.0306 ± 0.0092	0.8
(3)	0.0309 ± 0.0176	0.0300 ± 0.0106	0.0231 ± 0.0090	54.8
(4)	0.0098 ± 0.0022	**0.0124 ± 0.0017**	**0.0164 ± 0.0017**	2.7
	Space: Test accuracy - #MACs			
(1)	0.0195 ± 0.0131	0.0348 ± 0.0129	0.0250 ± 0.0069	53.8
(2)	0.0189 ± 0.0069	0.0322 ± 0.0085	0.0197 ± 0.0036	0.8
(3)	0.0266 ± 0.0150	0.0247 ± 0.0083	0.0188 ± 0.0060	54.8
(4)	**0.0104 ± 0.0023**	**0.0137 ± 0.0024**	0.0163 ± 0.0022	2.7

Table 4: IGD on transfer learning task: (1) MOENAS, (2) TF-MOENAS, (3) MaOENAS, (4) TF-MaOENAS. Results that are underlined indicate the best method and results that are **bolded** denote the best method with statistical significance (p-value < 0.01)

Alg.	CIFAR-10 (direct)	CIFAR-100 (transfer)	ImageNet16-120 (transfer)	Search cost (hours)
	Space: Test accuracy - FLOPs			
(1)	1.0332 ± 0.0013	0.7962 ± 0.0043	0.5167 ± 0.0051	53.7
(2)	1.0223 ± 0.0027	0.7830 ± 0.0054	0.5061 ± 0.0057	0.7
(3)	1.0334 ± 0.0011	0.7996 ± 0.0023	0.5191 ± 0.0028	54.8
(4)	1.0298 ± 0.0022	0.7958 ± 0.0039	0.5169 ± 0.0015	2.7
	Space: Test accuracy - Latency			
(1)	1.0050 ± 0.0006	0.7589 ± 0.0028	0.4861 ± 0.0056	54.1
(2)	0.9431 ± 0.0200	0.6425 ± 0.0284	0.4164 ± 0.0179	1.1
(3)	0.9967 ± 0.0168	0.7545 ± 0.0159	0.4897 ± 0.0067	54.8
(4)	0.9581 ± 0.0098	0.7234 ± 0.0140	0.4710 ± 0.0055	2.7
	Space: Test accuracy - #Parameters			
(1)	1.0332 ± 0.0014	0.7963 ± 0.0044	0.5155 ± 0.0055	53.8
(2)	1.0233 ± 0.0027	0.7824 ± 0.0054	0.5056 ± 0.0057	0.8
(3)	1.0334 ± 0.0011	**0.7995 ± 0.0023**	0.5189 ± 0.0028	54.8
(4)	1.0296 ± 0.0023	0.7954 ± 0.0040	0.5166 ± 0.0016	2.7
	Space: Test accuracy - #MACs			
(1)	1.0331 ± 0.0017	0.7964 ± 0.0048	0.5165 ± 0.0055	53.8
(2)	1.0280 ± 0.0034	0.7803 ± 0.0058	0.5042 ± 0.0060	0.8
(3)	1.0333 ± 0.0011	0.7992 ± 0.0023	0.5186 ± 0.0028	54.8
(4)	1.0292 ± 0.0025	0.7947 ± 0.0042	0.5160 ± 0.0016	2.7

Table 5: Hypervolume on transfer learning task: (1) MOENAS, (2) TF-MOENAS, (3) MaOENAS, (4) TF-MaOENAS. Results that are underlined indicate the best method and results that are **bolded** denote the best method with statistical significance (p-value < 0.01)

spaces simultaneously, rather than having to run separately TF-MOENAS many times for each bi-objective space.

We note that the variation in the obtained results across the datasets (see Tables 1, 2, 3) can be attributed to the following reasons. First, the performance metrics (i.e., accuracy or `synflow`) and some complexity metrics (e.g., latency or FLOPs) of each candidate architecture vary across the datasets (e.g., the accuracy of an architecture on CIFAR-10 is different from its accuracy on ImagetNet16-120). Therefore, the IGD and hypervolume results of each NAS method are different from one dataset to another. Second, we assess the effectiveness of NAS methods using the test accuracy after 200 training epochs but, during the search process of each NAS algorithm, the validation accuracy after 12 training epochs (for training-based approaches) or `synflow` (for training-free approaches) are employed as the performance objective (see experimental details in Section 4.1). The correlation of 12-epoch validation accuracy or `synflow` with the final test accuracy (after 200 epochs) varies per dataset [1] (e.g., the correlation coefficients of `synflow` for CIFAR-10, CIFAR-100, and ImageNet16-120 are 0.74, 0.76, and 0.75, respectively). Therefore, the rankings of the considered NAS methods might differ across the datasets.

4.4 Tranferability

This section explores the potential of transfer learning in NAS by evaluating the transferability of architectures discovered through multi-objective and many-objective NAS problem formulations. The final approximation front (i.e., the elitist archive) of architectures on CIFAR-10 is re-evaluated on CIFAR-100 and ImageNet16-120 for their performance and complexity. Transfer learning in NAS offers several benefits, including the reduced computational cost and the potential for faster deployment of deep learning models in real-world applications by identifying architectures that are highly transferable across datasets.

Table 4 and Table 5 show that TF-MaOENAS yields better IGD compared to other methods, whereas MaOENAS outperforms other methods in hypervolume in most cases. In terms of test accuracy, TF-MaOENAS also completely surpasses most of the approaches in Table 6, with better accuracy and lower search costs. It indicates that TF-MaOENAS using the training-free performance metric `synflow` are more effective at transferring knowledge from one dataset to another. Besides, both penta-objective approaches TF-MaOENAS and MaOENAS give better IGD and hypervolume, respectively, than bi-objective approaches. Although the four TF-MOENAS approaches have lower computing time, the optimization result of TF-MaOENAS is a penta-objective approximation front that

	CIFAR-10 (direct)	CIFAR-100 (transfer)	ImageNet16-120 (transfer)	Search cost (hours)
Manually designed				
ResNet [14]	93.97	70.86	43.63	-
Weight sharing				
RSPS [16]	87.66 ± 1.69	58.33 ± 4.34	31.14 ± 3.88	2.1
DARTS [17]	54.30 ± 0.00	15.61 ± 0.00	16.32 ± 0.00	3.0
GDAS [10]	93.51 ± 0.13	70.61 ± 0.26	41.84 ± 0.90	8.0
SETN [9]	86.19 ± 4.63	56.87 ± 7.77	31.90 ± 4.07	8.6
ENAS [24]	54.30 ± 0.00	15.61 ± 0.00	16.32 ± 0.00	3.6
Non-weight sharing				
RS [2]	93.70 ± 0.36	71.04 ± 1.07	44.57 ± 1.25	3.3
BOHB [13]	93.61 ± 0.52	70.85 ± 1.28	44.42 ± 1.49	3.3
NASWOT* [22]	93.84 ± 0.23	71.56 ± 0.78	45.67 ± 0.64	-
Evolution				
REA [27]	93.92 ± 0.30	71.84 ± 0.99	45.54 ± 1.03	3.3
TF-MOENAS*† [7]	94.16 ± 0.22	72.75 ± 0.63	46.61 ± 0.46	2.87
MOENAS (`valacc` - FLOPs)†	94.28 ± 0.17	72.68 ± 0.71	46.50 ± 0.68	53.7
TF-MOENAS (`synflow` - FLOPs)*†	94.29 ± 0.17	73.22 ± 0.71	46.31 ± 0.40	0.7
MOENAS (`valacc` - Latency)†	94.30 ± 0.09	73.00 ± 0.32	46.35 ± 0.43	54.1
TF-MOENAS (`synflow` - Latency)*†	94.29 ± 0.14	73.17 ± 0.25	46.28 ± 0.31	1.1
MOENAS (`valacc` - #Parameters)†	94.27 ± 0.18	72.72 ± 0.69	46.31 ± 0.68	53.8
TF-MOENAS (`synflow` - #Paramters)*†	94.24 ± 0.22	72.81 ± 0.76	46.31 ± 0.32	0.8
MOENAS (`valacc` - #MACs)†	94.24 ± 0.22	72.60 ± 0.77	46.37 ± 0.74	53.8
TF-MOENAS (`synflow` - #MACs)*†	94.35 ± 0.03	73.15 ± 0.07	46.47 ± 0.00	0.8
MaOENAS†	94.27 ± 0.13	72.94 ± 0.33	46.53 ± 0.27	54.8
TF-MaOENAS*†	94.37 ± 0.00	73.51 ± 0.00	46.51 ± 0.04	2.7
Optimal	**94.37**	**73.51**	**47.31**	-

* Training-Free †Multi-Objective/Many-Objective

Table 6: Accuracy on the transfer learning task. Previous studies' results are adopted from [11, 22]. Results that are underlined indicate the best method

contains much more insightful information, which can be obtained in one run and easily projected into any lower-dimensional objective spaces for intuitive Pareto front investigations.

5 Conclusions

This paper described different multi-objective and many-objective problem formulations for NAS, i.e., MONAS and MaONAS, which can be solved by multi-objective evolutionary algorithms, such as NSGA-II. We showed that the training-free metric `synflow` can be used as a proxy metric for the network accuracy performance during NAS, without requiring any training epochs. Experimental results demonstrated the benefits of using training-free approaches, especially the many-objective TF-MaOENAS, including computational efficiency, search effectiveness and insightful decision-making capabilities. These benefits were due to the ability to obtain top-performing architectures on both direct and transfer learning tasks, and the resulting penta-objective fronts of non-dominated architectures, which provided beneficial trade-off information among the concerned objectives.

References

[1] Mohamed S. Abdelfattah, Abhinav Mehrotra, Lukasz Dudziak, and Nicholas Donald Lane. 2021. Zero-Cost Proxies for Lightweight NAS. In *ICLR 2021*. OpenReview.net. `https://openreview.net/forum?id=0cmMMy8J5q`

[2] James Bergstra and Yoshua Bengio. 2012. Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.* 13 (2012), 281–305. `https://doi.org/10.5555/2503308.2188395`

[3] Peter A. N. Bosman and Dirk Thierens. 2003. The balance between proximity and diversity in multiobjective evolutionary algorithms. *IEEE Trans. Evol. Comput.* 7, 2 (2003), 174–188. `https://doi.org/10.1109/TEVC.2003.810761`

[4] Wuyang Chen, Xinyu Gong, and Zhangyang Wang. 2021. Neural Architecture Search on ImageNet in Four GPU Hours: A Theoretically Inspired Perspective. In *ICLR 2021*. OpenReview.net. `https://openreview.net/forum?id=Cnon5ezMHtu`

[5] Kalyanmoy Deb. 2001. *Multi-objective optimization using evolutionary algorithms*. Wiley.

[6] Kalyanmoy Deb, Samir Agrawal, Amrit Pratap, and T. Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* 6, 2 (2002), 182–197. `https://doi.org/10.1109/4235.996017`

[7] Tu Do and Ngoc Hoang Luong. 2021. Training-Free Multi-objective Evolutionary Neural Architecture Search via Neural Tangent Kernel and Number of Linear Regions. In *ICONIP 2021 (Lecture Notes in Computer Science, Vol. 13109)*, Teddy Mantoro, Minho Lee, Media Anugerah Ayu, Kok Wai Wong, and Achmad Nizar Hidayanto (Eds.). Springer, 335–347. `https://doi.org/10.1007/978-3-030-92270-2_29`

[8] Xuanyi Dong, Lu Liu, Katarzyna Musial, and Bogdan Gabrys. 2022. NATS-Bench: Benchmarking NAS Algorithms for Architecture Topology and Size. *IEEE*

Trans. Pattern Anal. Mach. Intell. 44, 7 (2022), 3634–3646. https://doi.org/10.1109/TPAMI.2021.3054824

[9] Xuanyi Dong and Yi Yang. 2019. One-Shot Neural Architecture Search via Self-Evaluated Template Network. In *ICCV 2019*. IEEE, 3680–3689. https://doi.org/10.1109/ICCV.2019.00378

[10] Xuanyi Dong and Yi Yang. 2019. Searching for a Robust Neural Architecture in Four GPU Hours. In *CVPR 2019*. Computer Vision Foundation / IEEE, 1761–1770. https://doi.org/10.1109/CVPR.2019.00186

[11] Xuanyi Dong and Yi Yang. 2020. NAS-Bench-201: Extending the Scope of Reproducible Neural Architecture Search. In *ICLR 2020*. OpenReview.net. https://openreview.net/forum?id=HJxyZkBKDr

[12] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. 2019. Neural Architecture Search: A Survey. *J. Mach. Learn. Res.* 20 (2019), 55:1–55:21. http://jmlr.org/papers/v20/18-598.html

[13] Stefan Falkner, Aaron Klein, and Frank Hutter. 2018. BOHB: Robust and Efficient Hyperparameter Optimization at Scale. In *ICML 2018 (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer G. Dy and Andreas Krause (Eds.). PMLR, 1436–1445. http://proceedings.mlr.press/v80/falkner18a.html

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR 2016*. IEEE Computer Society, 770–778. https://doi.org/10.1109/CVPR.2016.90

[15] Zhilin He. 2023. Improved Genetic Algorithm in Multi-objective Cargo Logistics Loading and Distribution. *Informatica* 47, 2 (2023). https://doi.org/10.31449/inf.v47i2.3958

[16] Liam Li and Ameet Talwalkar. 2019. Random Search and Reproducibility for Neural Architecture Search. In *UAI 2019 (Proceedings of Machine Learning Research, Vol. 115)*, Amir Globerson and Ricardo Silva (Eds.). AUAI Press, 367–377. http://proceedings.mlr.press/v115/li20c.html

[17] Hanxiao Liu, Karen Simonyan, and Yiming Yang. 2019. DARTS: Differentiable Architecture Search. In *ICLR 2019*. OpenReview.net. https://openreview.net/forum?id=S1eYHoC5FX

[18] Zhichao Lu, Ran Cheng, Yaochu Jin, Kay Chen Tan, and Kalyanmoy Deb. 2022. Neural Architecture Search as Multiobjective Optimization Benchmarks: Problem Formulation and Performance Assessment. *IEEE Transactions on Evolutionary Computation* (2022). https://doi.org/10.1109/TEVC.2022.3233364

[19] Zhichao Lu, Kalyanmoy Deb, Erik D. Goodman, Wolfgang Banzhaf, and Vishnu Naresh Boddeti. 2020. NSGANetV2: Evolutionary Multi-objective Surrogate-Assisted Neural Architecture Search. In *ECCV 2020 (Lecture Notes in Computer Science, Vol. 12346)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer, 35–51. https://doi.org/10.1007/978-3-030-58452-8_3

[20] Zhichao Lu, Ian Whalen, Yashesh D. Dhebar, Kalyanmoy Deb, Erik D. Goodman, Wolfgang Banzhaf, and Vishnu Naresh Boddeti. 2020. NSGA-Net: Neural Architecture Search using Multi-Objective Genetic Algorithm (Extended Abstract). In *IJCAI 2020*, Christian Bessiere (Ed.). ijcai.org, 4750–4754. https://doi.org/10.24963/ijcai.2020/659

[21] Hoang N. Luong and Peter A. N. Bosman. 2012. Elitist Archiving for Multi-Objective Evolutionary Algorithms: To Adapt or Not to Adapt. In *PPSN XII (Lecture Notes in Computer Science, Vol. 7492)*, Carlos A. Coello Coello, Vincenzo Cutello, Kalyanmoy Deb, Stephanie Forrest, Giuseppe Nicosia, and Mario Pavone (Eds.). Springer, 72–81. https://doi.org/10.1007/978-3-642-32964-7_8

[22] Joseph Charles Mellor, Jack Turner, Amos J. Storkey, and Elliot J. Crowley. 2020. Neural Architecture Search without Training. *CoRR* abs/2006.04647 (2020). arXiv:2006.04647 https://arxiv.org/abs/2006.04647

[23] Sarat Mishra and Sudhansu Kumar Mishra. 2020. Performance Assessment of a set of Multi-Objective Optimization Algorithms for Solution of Economic Emission Dispatch Problem. *Informatica* 44, 3 (2020), 349–360. https://doi.org/10.31449/inf.v44i3.1969

[24] Hieu Pham, Melody Y. Guan, Barret Zoph, Quoc V. Le, and Jeff Dean. 2018. Efficient Neural Architecture Search via Parameter Sharing. In *ICML 2018 (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer G. Dy and Andreas Krause (Eds.). PMLR, 4092–4101. http://proceedings.mlr.press/v80/pham18a.html

[25] Quan Minh Phan and Ngoc Hoang Luong. 2021. Efficiency Enhancement of Evolutionary Neural Architecture Search via Training-Free Initialization. In *(NICS) 2021*. 138–143. https://doi.org/10.1109/NICS54270.2021.9701573

[26] Quan Minh Phan and Ngoc Hoang Luong. 2023. Enhancing multi-objective evolutionary neural architecture search with training-free Pareto local search.

Appl. Intell. 53, 8 (2023), 8654–8672. https://doi.org/10.1007/s10489-022-04032-y

[27] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V. Le. 2019. Regularized Evolution for Image Classifier Architecture Search. In *AAAI 2019*. AAAI Press, 4780–4789. https://doi.org/10.1609/aaai.v33i01.33014780

[28] Hidenori Tanaka, Daniel Kunin, Daniel L. K. Yamins, and Surya Ganguli. 2020. Pruning neural networks without any data by iteratively conserving synaptic flow. In *NeurIPS 2020*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/46a4378f835dc8040c8057beb6a2da52-Abstract.html

[29] An Vo, Tan Ngoc Pham, Van Bich Nguyen, and Ngoc Hoang Luong. 2022. Training-Free Multi-Objective and Many-Objective Evolutionary Neural Architecture Search with Synaptic Flow. In *SoICT 2022*. ACM, 1–8. https://doi.org/10.1145/3568562.3568569

[30] Barret Zoph and Quoc V. Le. 2017. Neural Architecture Search with Reinforcement Learning. In *ICLR 2017*. OpenReview.net. https://openreview.net/forum?id=r1Ue8Hcxg

An Automatic Labeling Method for Subword-Phrase Recognition in Effective Text Classification

Yusuke Kimura[1], Takahiro Komamizu[2] and Kenji Hatano[3]
[1]Graduate School of Culture and Information Science, Doshisha University, Japan
[2]Mathematical and Data Science Center, Nagoya University, Japan
[3]Faculty of Culture and Information Science, Doshisha University, Japan
E-mail: usk@acm.org, taka-coma@acm.org, hatano@acm.org

Keywords: text classification, subword-phrase, multi-task learning

Received: March 16, 2023

The deep learning-based text classification methods perform better than traditional ones. In addition to the success of the deep learning technique, multi-task learning (MTL) has come to become a promising approach for text classification; for instance, an MTL approach in text classification employs named entity recognition as an auxiliary task and has showcased that the task helps to improve the text classification performance. Existing MTL-based text classification methods depend on the auxiliary tasks using supervised labels. Obtaining such supervision labels requires additional human and financial costs in addition to those for the main text classification task. To reduce these additional costs, we propose an MTL-based text classification framework on supervised label creation by automatically labeling phrases in texts for the auxiliary recognition task. A basic idea to realize the proposed framework is to utilize phrasal expressions consisting of subwords (called subword-phrases). To the best of our knowledge, no text classification approach has been designed on top of subword-phrases because subwords only sometimes express a coherent set of meanings. The novelty of the proposed framework is in adding subword-phrase recognition as an auxiliary task and utilizing subword-phrases for text classification. It extracts subword-phrases in an unsupervised manner using the statistics approach. To construct labels for effective subword-phrase recognition tasks, extracted subword-phrases are classified based on document classes to ensure that subword-phrases dedicated to some classes can be distinguishable. Experimental evaluation for text classification using five popular datasets showcased the effectiveness of the subword-phrase recognition as an auxiliary task. It also showed that comparing various labeling schemes in recent studies indicated insights for labeling common subword-phrases among several document classes.

Povzetek: Za klasifikacijo besedil je uporabljeno globoko učenje in večopravilno učenje iz uporabo podbesednih fraz za avtomatsko označevanje.

1 Introduction

Text classification is a fundamental technology that has been studied for a long time. Applications that use text classification include speech [7], categorizing daily news articles, and unfair clause detection in terms of services [15]. These text classification applications are achieved by effectively and efficiently retrieving information from large amounts of text [12, 23]. Text classification is a supervised learning task manually assigning labels to documents as classification criteria, such as categories and classes. A classifier learns classification criteria in a feature space based on the dataset. Traditionally, text classification uses hand-crafted features such as term frequency-inverse document frequency. In recent literature, deep learning-based technologies have achieved significantly improved classification performance. A component that has improved text classification performance in recent years is pre-trained neural language models such as BERT, which have been trained on vast amounts of text. Pre-trained neural language models provide semantically rich features for text; therefore, even a simple multi-layer perceptron-based classifier has performs excellently. After the initial success of BERT, many pre-trained models, such as RoBERTa [19] and GPT-3 [5], have been published.

The tokenizers in these pre-trained neural language models typically divide documents into subwords as the smallest unit. Subwords reduce the number of unknown words not in the vocabulary, thus preventing the performance of pre-trained neural language models from being degraded by unknown words. Subword-based tokenization effectively handles out-of-vocabulary (OOV) words by decomposing such words into several subwords. Concatenations of these subwords represent OOV words, while traditional approaches represent them as *unknown* tokens. The subword-based tokenization was initially employed for machine translation [29]; after that, it was used in various natural language processing tasks, including text classification.

Multi-task learning (MTL) [6, 37, 39], which involves one or more auxiliary tasks with the primary task by sharing parameters, is a promising approach to enhance the performance of deep learning models. It has also been applied to text classification [17,35,36]. Learning models with auxiliary tasks positively affect the generalization performance of the main task and reduce over-fitting. Early studies on MTL-based text classification [17,35] focused on methods to combine multiple tasks and combined tasks in different datasets. Recent studies have combined text classification with auxiliary tasks using the same dataset, such as named entity recognition (NER) [2,31] or label co-occurrence prediction [36].

The fact that MTL with NER and text classification improves the accuracy of text classification performance suggests that the recognition of clause representations, such as named entities, is suitable as an auxiliary task to MTL-based text classification. However, to realize NER as an auxiliary task for MTL-based text classification, supervised labels for NER are required in addition to those for text classification. Constructing such training datasets is costly because of additional human costs for NER labeling.

Therefore, in this study, we seek to achieve MTL-based text classification with phrasal expression recognition, which does not require additional human cost to construct a training dataset. Phrasal expressions (or key phrases) for texts have been studied in past decades [27,38]. Applying keyphrase extraction based on the subword-based tokenization of popular pre-trained neural language models is not straightforward. Therefore, we define a phrasal expression based on subwords as a *subword-phrase* and seek its potential usability for the MTL-based text classification. In contrast to phrasal expressions based on words, subword-phrases are not necessarily semantically coherent because a vocabulary of subwords is determined statistically [29]. Owing to such little semantic coherence of subword-phrases, studies have never been conducted on their utilization for text classification.

In this study, we propose a framework for MTL-based text classification with subword-phrase recognition to improve the accuracy of text classification. Our framework comprises unsupervised subword-phrase labeling and MTL-based text classification for the subword-phrase recognition task. Notably, we assume the presence of labels for the classification of a dataset. To implement our framework, we employ a highly primitive approach: frequency-based subword-phrase labeling, in which frequently co-occurring consecutive subwords are merged to form a subword-phrase; various implementations can be realized using this approach. We also employ the concept of byte-pair encoding [29]. We seek labeling schemes to handle commonly appearing subword-phrases among document classes to make the auxiliary task more effective than text classification tasks.

The contributions of this study can be summarized as follows: MTL-based text classification with low-cost auxiliary task preparation, utilization of phrasal expression for subwords, and superior performance over conventional methods, and comparable performance with the novel methods. The proposed framework comprises an unsupervised labeling module and an MTL-based classification module. Existing MTL-based text classification methods assume the presence of supervision for auxiliary tasks; however, obtaining this supervision requires further human and financial costs. In contrast, the proposed framework does not require these costs as it utilizes unsupervised subword-phrase extraction to obtain labels to create auxiliary tasks.

Our method is the first study that utilizes subword-phrases. As subwords are not necessarily semantically coherent, their phrasal expressions have yet to be considered for any task. In contrast, the co-occurrence of consecutive subwords or subword-phrases could contribute to the text classification task. Such subwords may represent distinguished instances of a class from those of others. In the experimental evaluation of five popular text classification datasets, the proposed framework with subword-phrase recognition auxiliary task demonstrated improved classification performance (micro and macro F-scores) compared to the single-task method. Compared with the state-of-the-art method (BertGCN [14]), the proposed framework also demonstrated superior performance for datasets with more labels, exhibiting comparative classification performance for the other datasets.

The rest of this paper is organized as follows. Section 2 introduces studies concerning MTL-based text classification. Section 3 explains the proposed framework of MTL-based text classification with subword-phrase recognition task. Section 4 then presents the experimental evaluation, which demonstrates the effectiveness of the proposed framework compared to that of the single-task text classification baseline as well as other novel methods; it also discusses the effect of subword-phrases. Finally, Section 5 concludes this paper.

2 Related work

This section introduces literature related to MTL-based text classification. MTL-based text classification methods are categorized into the following three types based on the relationships between the main and auxiliary tasks [35]; Multi-Cardinality, Multi-Domain, and Multi-Objective.

Multi-Cardinality means that the main and auxiliary tasks are of different datasets but are in the same domain; these tasks also differ in cardinality, meaning that they vary in terms of their text lengths and the number of classes, among other parameters.

Multi-Domain means that the main and auxiliary tasks are similar, but their domains differ. For example, Liu et al. [16] and Zhang et al. [35] examined MTL-based movie review classification with classification tasks of reviews for various products, such as books and DVDs [4].

Multi-Objective means that the main and auxiliary tasks have different objectives. For example, Liu et al. [18] combined query classification and search result ranking using an MTL approach, and Zhang et al. [35] attempted MTL-based movie review classification (IMDB [21]) with news article classification (RN [1]) and question type classification (QC [13]) as auxiliary tasks.

In addition, MTL approaches [3, 30, 33, 40] in which the main and auxiliary tasks are in the same dataset have exhibited their effectiveness. Bi et al. [3] improved the performance of news recommendations by using MTL, which combines the news recommendation task with news article classification and named entity recognition. The MTL-based medical query intent classification model, proposed by Tohti et al. [30], was trained together with the named entity recognition, and consequently showed superior classification performance. On another task, Yang et al. [33] and Zhao et al. [40] showed similar observations on polarity classification combined with the aspect term extraction task. In the emotion prediction task, Li et al. [11] dealt with the emotion-cause pair extraction task using the MTL-based approach, which is combined with the emotion clause extraction and the cause clause extraction. Similarly, Qi et al. [24] proposed the MTL-based aspect sentiment classification method, where the auxiliary task was the aspect term extraction; they also demonstrated its effectiveness. In addition to the text classification task, the MTL-based approaches to image classification tasks have also shown its effectiveness [9, 32].

MTL-based text classification, which utilizes the relationship between labels in the same dataset, has also been proposed to solve the multi-label classification problem, where a single text can be classified into multiple labels [36]. Zhang et al. [36] showed improved classification performance by designing an auxiliary task to learn the relationship between labels.

These studies have shown the effectiveness of combining multiple supervised learning. However, in general, creating supervised data is expensive in terms of human and financial costs; thus, lower-cost solutions to design auxiliary tasks are desirable.

Self-supervised learning (SSL) is a training approach that understands data without supervised datasets. It first hides pieces of data and trains the model so that the model can estimate the hidden pieces. Masked language model (MLM) is a popular SSL in the natural language processing domain [8]. A popular pre-trained neural language model, BERT [8], is trained based on two SSL tasks: MLM and next sentence prediction. In the image processing domain, DALL-E [25] showcased the significant performance of SSL, where an area of an image was erased and DALL-E was trained to estimate the erased area. The increasing attention to these models indicates the usefulness of SSL for *data understanding* and *representation learning*.

In contrast to data understanding, text classification is a supervised learning task. In other words, SSL expects models to reconstruct broken pieces of data, while supervised learning expects models to learn dedicated criteria from supervision. Therefore, task settings in SSL are not easily imported to MTL-based text classification.

The proposed framework in this study focuses on creating datasets for auxiliary tasks with no supervision, significantly reducing human efforts and financial costs. To our knowledge, no research has been conducted that aimed to design auxiliary tasks of MTL-based text classification with no supervision. In addition, as subwords are not necessarily semantically coherent, subword-phrases have not been considered for any task. Therefore, this study proposes a novel methodology of MTL-based text classification in two aspects: In addition, since subwords are not necessarily semantically coherent, subword-phrases have not been considered for any task. Therefore, this paper proposes a novel methodology of MTL-based text classification in two aspects: (1) low-cost auxiliary task design and (2) introduction of subword-phrases. The experimental evaluation of this study reveals promising results for these two aspects.

3 Proposed framework

This section explains our framework of the MTL-based text classification, which generates subword-phrase labels for auxiliary tasks in an unsupervised manner.

3.1 Framework overview

Figure 1 illustrates our framework. It consists of two phases: unsupervised labeling and MTL-based text classification. The basic approach underlying of the framework is that subword-phrase recognition is added as an auxiliary task for MTL-based text classification. To realize the recognition task, unsupervised subword-phrase extraction is employed to create pseudo-supervision. A text classifier based on the framework is trained using the following steps:

1. **Input**: the text classifier receives a training set of text with classification labels;
2. **Tokenization**: the text is tokenized into subwords using a subword-based tokenizer;
3. **Labeling (Phase 1)**: the unsupervised labeling module appends subword-phrase labels to each text in the training set for the auxiliary subword-phrase recognition task;
4. **Training (Phase 2)**: the text classifier is trained in an MTL manner, which is trained together with the auxiliary subword-phrase recognition task based on the appended labels.

Formally, a training set is denoted as $D = \{(T_i, y_i) \mid 1 \leq i \leq N\}$, where T_i represents a sequence of subword tokens of the i-th text, y_i represents the class label corresponding to the i-th text, and N is the number of texts. In the first phase, the unsupervised subword-phrase labeling module receives D and performs subword-phrase extraction on subword token sequences to create another training set $D^{aug} = \{(T_i, Y_i^{aug}) \mid 1 \leq i \leq N\}$ for the auxiliary

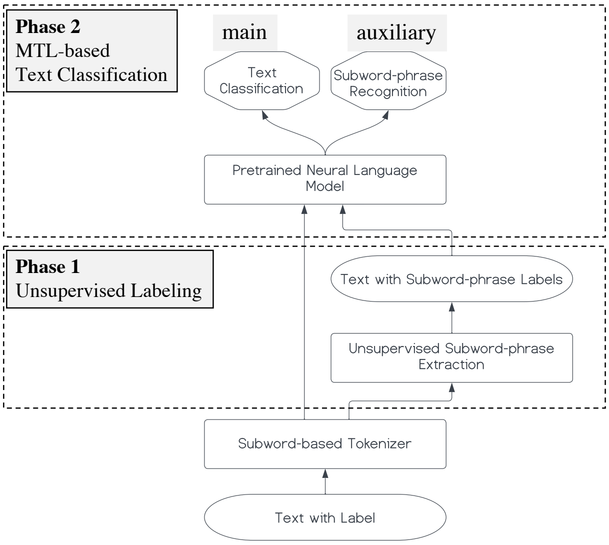


Figure 1: Our MTL-based Text Classification Framework. The framework accepts text with text classification labels and trains an MTL-based text classification model. The framework consists of two phases: the first phase is unsupervised labeling of the input text, and the second phase is the training of the MTL-based text classification model using the text classification labels and labels from the first phase.

task, where Y_i^{aug} is a corresponding sequence of labels for each token in T_i. In the second phase, D and D^{aug} are passed to an MTL-based text classification module based on a pre-trained neural language model; they then train the text classification model in conjunction with the training subword-phrase recognition model.

3.2 Unsupervised subword-phrase labeling

Unsupervised subword-phrase labeling provides a label sequence that corresponds to the input text sequence. This unsupervised labeling is a task formalized as follows:

- **Given**: a sequence of subword tokens T along with a class label y, $(T, y) \in D$
- **Generate**: a sequence of labels Y^{aug} whose length is exactly the same as that of T

The labeling scheme is inspired by NER tasks that employ the inside-outside-beginning (IOB2) tagging scheme [26]. IOB2 tagging is a labeling scheme where the first token of a phrase is tagged with `B` (beginning), the intermediate tokens of a phrase are tagged with `I` (inside), and tokens other than the phrase are tagged with `O` (outside). Besides these tags, semantic types are appended to distinguish types of phrases; for example, `B-PERSON` and `I-PERSON` represent the beginning and intermediate tokens of a token sequence corresponding with a person's name, respectively.

A straightforward labeling scheme for subword-phrase labeling is to treat all phrases equally. In other words, the semantic type is set to `Phrase`. Formally, when an n-length sequence of tokens $S = (s_1, s_2, \ldots, s_n)$ has a phrase which is an m-length sub-sequence $P = (s_k, s_{k+1}, \ldots, s_{k+m})$ of S where $m \leq n$, s_k is labeled as a particular type `B-Phrase`; the rest of the tokens from s_{k+1} to s_{k+m} are labeled as `I-Phrase` and other tokens $s_i \in S \backslash P$ are labeled as `O`.

This approach is so straightforward that subword-phrases appearing in different document classes are treated equally. However, to provide cues to the main text classification model, subword-phrases dependent on document classes should be distinguishable. A simple classification-specific labeling scheme assigns different labels to subword-phrases appearing in other classes. When a subword-phrase $P = (s_k, s_{k+1}, \ldots, s_{k+m})$, which is a sequence of tokens of a text belonging to class y, s_k is labeled as `B-`y, and the remaining tokens from s_{k+1} to s_{k+m} are labeled as `I-`y. However, subword-phrases commonly appearing in different classes cannot be handled in this scheme. To handle such common subword-phrases, we propose three labeling schemes, namely, **Disregard**, **Common-Label**, and **Bit-Label**. To compare, the aforementioned straightforward labelling scheme is called **All-Phrase**. Disregard scheme simply ignores the common subword-phrases, in other words, they are labeled by `O` tags. In Common-Label scheme, a special class label $\emptyset$ is used as a special semantic type of labeling in the IOB2 scheme. Specifically, the common subword-phrase P is labeled as `B-`$\emptyset$ for s_k and `I-`$\emptyset$ for other tokens. To handle such subword-phrases, this study proposes a bit-encoding-based labeling scheme. Bit-Label scheme still inherits the IOB2 labeling scheme; therefore, suppose that $d = 4$, a subword-phrase $P = (s_k, s_{k+1}, \ldots, s_{k+m})$, which is a sequence of tokens of a text and belongs to the first and third classes, then s_k is labeled as `B-1010`, and the rest of the tokens from s_{k+1} to s_{k+m} are labeled as `I-1010`.

3.3 MTL-based text classification

Our framework uses a text classification model based on MTL and a pre-trained neural language model (NLM). In this method, the NLM performs token encoding, and classification modules for main and auxiliary tasks are appended on top of the encoding. Therefore, NLM is the part shared among tasks and is trained in an MTL manner. A fully connected layer and a softmax non-linear layer design the classification models.

For the main task (i.e., text classification), a representation $\mathbf{h}^{cls}$ for a given input token sequence is obtained from NLM. It is passed to a fully connected layer followed by a softmax layer to predict class distribution $\hat{\mathbf{y}}^{cls}$. Formally, $\hat{\mathbf{y}}^{cls}$ for $\mathbf{h}^{cls}$ is calculated by the following equation:

$$\hat{\mathbf{y}}^{cls} = \mathrm{softmax}(W_{cls}^{\top} \cdot \mathbf{h}^{cls} + \mathbf{b}^{cls}), \qquad (1)$$

where W_{cls} and $\mathbf{b}^{cls}$ denote the parameter matrix and bias, respectively, for the text classification task.

For the auxiliary tasks (i.e., subword-phrase recognition), a representation $\mathbf{h}_j^{spr}$ for the j-th token of a given input sequence is obtained from NLM. It is passed to a fully connected layer followed by a softmax layer to predict token label distribution $\hat{\mathbf{y}}^{spr}$. Formally, $\hat{\mathbf{y}}_j^{spr}$ for $\mathbf{h}_j^{spr}$ is calculated by the following equation:

$$\hat{\mathbf{y}}_j^{spr} = \mathrm{softmax}(W_{spr}^{\top} \cdot \mathbf{h}_j^{spr} + \mathbf{b}^{spr}), \tag{2}$$

where W_{spr} and $\mathbf{b}^{spr}$ denote the parameter matrix and bias, respectively, for the subword-phrase recognition task.

These main and auxiliary tasks are multi-class classification tasks; therefore, using the cross-entropy loss as a loss function is straightforward. The following equation calculates the loss L_{cls} for the text classification task:

$$L_{cls} = -\sum_{i=1}^{N} \sum_{c \in C} y_{i,c} \log \hat{y}_{i,c}^{cls}, \tag{3}$$

where N is the number of training sample texts, C denotes a set of classes, $y_{i,c} \in \{0, 1\}$ denotes a true label for the i-th text where $y_{i,c} = 1$ if the true label of the text is c and 0 otherwise, and $\hat{y}_{i,c}^{cls}$ denotes the predicted probability of class c for the text.

Similarly, the following equation calculates the loss L_{spr} for the subword-phrase recognition task:

$$L_{spr} = -\sum_{i=1}^{N} \sum_{j=1}^{M_i} \sum_{c \in C} y_{i,j,c} \log \hat{y}_{i,j,c}^{cls}, \tag{4}$$

where N denotes the number of training sample texts, M_i denotes the number of tokens in the i-th text, C denotes a set of classes, $y_{i,j,c} \in \{0, 1\}$ denotes a true label for the j-th token of the i-th text where $y_{i,j,c} = 1$ if the true label of the token is c and 0 otherwise, and $\hat{y}_{i,j,c}^{spr}$ denotes the predicted probability of class c for that token.

To train both tasks simultaneously, feedback from results on these tasks is fed to the NLM model to fine-tune its parameters. Therefore, joint loss L_{joint} of losses for these tasks are calculated using the following equation and used for parameter optimization.

$$L_{joint} = L_{cls} + L_{spr} \tag{5}$$

We note that the weighting scheme in MTL approaches to involve the importance of individual tasks has been studied [22, 28]. Although considering the weighting scheme in our framework is promising, the purpose of this study is to show the capability of MTL-based text classification in conjunction with subword-phrase recognition, whoselabels for auxiliary tasks are created in an unsupervised manner. Therefore, employing the weighting scheme in our framework can be the focus of future studies.

4 Experimental evaluation

To evaluate the proposed framework, we conducted an experimental evaluation to answer the following items: (1) Whether or not our MTL-based text classification methods that create auxiliary tasks in an unsupervised manner improve classification performance compared to single-task text classification methods?, (2) Whether or not our MTL-based text classification can outperform state-of-the-art (SOTA) text classification methods?, (3) Whether or not the subword-phrase technique contributes to text classification?, and (4) Whether or not there is the best labeling scheme for subword-phrase recognition in terms of common subword-phrases?

The rest of this section is organized as follows: Section 4.1 introduces the implementation of the proposed framework; Section 4.2 explains the SOTA text classification method for comparison; Section 4.3 describes the experimental settings; Section 4.4 showcases the experimental results, and Section 4.5 presents remarks on the experiments by answering items mentioned above.

4.1 Implementation of the proposed framework

In this experiment, we implemented a simple frequency-based subword-phrase extraction method; the labeling scheme used for the extracted subword-phrase was the classification-specific labeling scheme. The frequency-based method expects that frequently co-occurring subwords compose the regular textual expressions for each class. To control the number of subword-phrases, we utilized the byte-pair encoding (BPE) algorithm [29]. The BPE algorithm concatenates consecutive tokens if they frequently co-occur in a corpus and repeats this concatenation until the number of unique tokens equals the expected number. The ability to control the number of subword-phrases was suitable for this experiment because the subword-phrase was newly proposed in this study; therefore, we needed to try variations of evaluation experiments which were realized by creating different numbers of subword-phrases.

In general, the number of texts is skewed among classes; the number of particular texts of a class may be quite large, while that of other classes is very small. This affected the extraction of subword-phrases; therefore, in this experiment, the extraction mentioned above was applied for each set of texts of class. Specifically, we extracted n subword-phrases for each class. n was chosen from $\{10, 100, 1000, 10000\}$ to achieve the best classification performance on the validation data.

4.2 Comparison method: BertGCN

BertGCN [14] is a SOTA method for text classification that combines a pre-trained NLM with the inductive learning of graph neural networks (GNNs). BertGCN follows TextGCN [34] by constructing a graph of the cooccurrence relations between texts and words and between words and words. In BertGCN, vectors of vertices are initialized using the pre-trained NLM. These vectors are updated through graph convolutional neural network (GCN) to involve the co-occurrence relationships between texts and words. Based on the updated vectors, BertGCN performs text classification by adding a fully connected layer followed by a softmax layer. In addition, [14] reported that integrating the output of the NLM-based classification model and that of BertGCN can improve classification performance; specifically, the linear sum of the predicted class distributions Z_{GCN} and Z_{NLM}, which are obtained from BertGCN and the classifier using NLM, respectively, as seen in the following equation:

$$Z = \lambda \cdot Z_{\text{GCN}} + (1 - \lambda) \cdot Z_{\text{NLM}}, \tag{6}$$

where $\lambda \in [0, 1]$ denotes the weight for BertGCN classification. This experiment used $\lambda = 0.7$ as [14] reported that it was the optimal value. BertGCN can use any pre-trained NLM, and [14] reported that RoBERTa showed the optimal performance. Therefore, RoBERTa was also used in implementing the proposed framework to make the comparison as reasonable as possible.

4.3 Settings

Datasets For the evaluation, the following five popular datasets in the text classification task are used; Movie Review (MR), 20 Newsgroups (20NG), R8, R52 and Ohsumed (OHS). MR is a dataset of movie reviews categorized into binary sentiment classes (i.e., positive and negative). 20NG is a dataset of news texts categorized into 20 categories. R8 is a dataset of news articles from Reuters-21578[1] limited to eight selected classes. R52 is a dataset of news articles from Reuters-21578 limited to 52 selected categories. OHS is a dataset of medical abstracts categorized into 23 medical concepts called MESH categories.

The statistics of the dataset are shown in Table 1. As the table shows, datasets with different classes and variations in the number of instances per class (the standard deviation (Std.) of the number of instances within a class) were used in the experiment. These datasets were expected to reveal the advantages and disadvantages of the proposed method.

Metrics The evaluation metric is F-score which is the harmonic mean of precision and recall scores as shown below.

$$\begin{aligned} Pre &= \frac{TP}{TP + FP} && (7) \\ Rec &= \frac{TP}{TP + FN} && (8) \\ F &= \frac{2 \cdot Pre \cdot Rec}{Prec + Rec} && (9) \end{aligned}$$

The precision, denoted by *Pre* is the ratio of the number of true positives (TP) over the number of instances estimated as positive (i.e., $TP + FP$, where FP is the number of false positives). The recall, denoted by *Rec* is the ratio of TP over the number of positive instances in the evaluation set (i.e., $TP + FN$, where FN is the number of false negatives). To observe various aspects for evaluation, micro and macro averages of F-scores were used in this experiment. The micro average of F-scores, F_{micro}, is the instance-level average of the F-score, and the macro average, F_{macro}, is the class-level average of the F-scores. When the numbers of instances of different classes are highly skewed (class imbalance problem), the F_{micro} is not suitable to evaluate the classification performance; this is because the larger the number of instances of a class, the more it affects this metric. In other words, the classification performance in the instances of minority classes is underestimated. In contrast, the F_{macro} metric can ignore the skewness as the F scores of difference classes are treated independently and averaged.

Parameters For the base model in the proposed method and BertGCN, we employed the RoBERTa-base model [19], available at Huggingface[2]. BertGCN with the RoBERTa model was called RoBERTaGCN in this experiment. In this study, the effect of common subword-phrases was also evaluated; therefore, the proposed method had two variations: one included common subword-phrases (denoted as Proposed w/ cmn) and the other excluded them (denoted as Proposed w/o cmn). In addition, as a baseline method, we also employed a single-task text classification method based on RoBERTa. The baseline method was implemented by adding a fully connected layer and a softmax layer on top of RoBERTa, which is equivalent to Eq. 1 with the loss function shown in Eq. 3. The only difference between the proposed and the baseline methods was the number of tasks on top of RoBERTa. Therefore, the comparison between them was expected to reveal the effectiveness of MTL-based text classification. These models were optimized using the AdamW optimizer (Adam optimizer [10] with decoupled weight decay regularization) [20]. Experiments were conducted with 100 epochs, batch size 64, and a maximum token length of 256. Only the experiment for RoBERTaGCN was conducted with a batch size of 128 and a maximum token length of 128, which yielded better results than the aforementioned hyper parameters.

4.4 Results

Table 2 shows the experimental results of F_{micro} (Table 2(a)) and F_{macro} (Table 2(b)), and showcases the following three observations. (1) The proposed method performed better than the baseline method in both metrics except the simple binary classification on the MR dataset. (2) The proposed method outperformed RoBERTaGCN for three of the five datasets in terms of the F_{micro} metric and four of the five datasets in terms of the F_{macro} metric. (3) In terms of labeling schemes, the Bit-Label and the Disregard approaches

[1] Reuters-21578, `http://www.daviddlewis.com/resources/testcollections/reuters21578/`, visited on Aug. 4, 2022

[2] `https://huggingface.co/roberta-base`

Table 1: Statistics of datasets. The number of instances in train-valid-test splits, number of classes, and average (Avg.) and standard deviation (Std.) of the number of instances across classes.

	MR	20NG	R8	R52	OHS
#Train	6,398	10,183	4,937	5,879	3,022
#Valid	710	1,131	548	653	335
#Test	3,554	7,532	2,189	2,568	4,043
#Class	2	20	8	52	23
Avg. #Instances/Class	5,331	942	959	175	321
Std. #Instances/Class	0	94	1,309	613	305

Table 2: Evaluation results. The best score in each column (i.e., dataset) is bold-faced. RoBERTaGCN is the SOTA text classification method and Baseline is the single-task text classification based on the RoBERTa model. The proposed method has two variations: one, denoted as Proposed w/ cmn, includes common subword-phrases in the labeling scheme, and the other, denoted as Proposed w/o cmn, excludes them. (a) and (b) showcase the results of F_{micro} and F_{macro}, respectively.

(a) F_{micro}

Model	MR	20NG	R8	R52	OHS
RoBERTaGCN	0.880	**0.894**	**0.979**	0.944	**0.736**
Baseline (RoBERTa)	0.881	0.831	0.977	0.962	0.690
Proposed - All-Phrase	**0.888**	0.838	**0.979**	0.967	0.705
Proposed - Common-Label	0.860	0.850	0.978	0.967	0.704
Proposed - Bit-Label	0.882	0.846	**0.979**	0.968	0.711
Proposed - Disregard	0.866	0.851	**0.979**	**0.969**	0.711

(b) F_{macro}

Model	MR	20NG	R8	R52	OHS
RoBERTaGCN	0.880	**0.861**	0.925	0.756	0.605
Baseline (RoBERTa)	0.881	0.825	0.943	0.836	0.594
Proposed - All-Phrase	**0.888**	0.832	0.948	0.842	0.622
Proposed - Common-Label	0.860	0.845	0.947	0.841	0.610
Proposed - Bit-Label	0.882	0.840	0.953	**0.866**	0.636
Proposed - Disregard	0.866	0.845	**0.955**	0.851	**0.637**

performed better than other schemes in terms of the F_{macro} metric.

The comparison between the proposed method and the baseline method in both F_{micro} and F_{macro} revealed the effectiveness of the MTL-based approach, in which the auxiliary task was systematically constructed. In addition to insights from existing literature that MTL-based approaches using auxiliary tasks with supervision are effective, this experiment showcased the effectiveness of an MTL approach in which training data for an auxiliary task was generated in an unsupervised manner. The results showcase that low-cost auxiliary tasks for MTL-based text classification now demonstrate promising performance.

While the results of MR and R8 datasets showed comparable performances between the proposed and the baseline methods, these datasets were composed of smaller numbers of classes. These results suggest that the proposed method did not perform effectively when the number of classes was small.

A notable fact from the results was the proposed method achieved significantly better performance than RoBERTaGCN in terms of F_{macro} on the R8, R52, and OHS datasets. Simultaneously, the proposed method was also more accurate than RoBERTaGCN in terms of F_{micro}. These facts indicate that the proposed method achieved state-of-the-art classification performance on these datasets. Recalling the statistics of these datasets from Table 1, the numbers of classes in each R8, R52 and OHS dataset are larger than those of other datasets and the number of instances per class is highly skewed. These facts indicate that the proposed method is good for highly skewed datasets. Though 20NG dataset had similar number of classes to the OHS dataset and was less skewed than the OHS dataset, the performance in terms of F_{micro} and F_{macro} of the proposed

Table 3: Evaluation results: Accuracy of auxiliary tasks

(a) F_{micro}

Model	MR	20NG	R8	R52	OHS
Proposed - All-Phrase	**0.971**	**0.975**	**0.978**	**0.998**	0.971
Proposed - Common-Label	0.922	0.968	0.974	0.972	**0.978**
Proposed - Bit-Label	0.918	0.974	0.965	0.975	0.977
Proposed - Disregard	0.922	0.851	0.962	0.975	**0.978**

(b) F_{macro}

Model	MR	20NG	R8	R52	OHS
Proposed - All-Phrase	**0.960**	**0.975**	**0.945**	0.796	**0.953**
Proposed - Common-Label	0.761	0.889	0.869	0.853	0.725
Proposed - Bit-Label	0.756	0.852	0.764	**0.864**	0.762
Proposed - Disregard	0.761	0.845	0.731	0.847	0.725

method was worse than RoBERTaGCN. Consequently, the proposed method performed better than the SOTA method when datasets were composed of large classes and highly skewed in the number of instances across classes.

The comparison among variations of the proposed method in terms of the labeling schemes for commonly appearing subword-phrases among document classes showed that the proposed method with different schemes had similar performances, each with their pros and cons for different datasets. The All-Phrase scheme had all phrases labeled by the IOB2 tagging scheme regardless of document classes. Compared with other schemes that take document classes into account, its performance was inferior. This indicates that class-specific labeling (the Common-Label, Bit-Label, and Disregard schemes) is effective, except for the MR dataset, which is a binary classification dataset; thus, subword-phrases are merely *class-specific*. For the comparison of labeling common subword-phrases among the Common-Label, Bit-Label, and Disregard schemes, their classification performances were comparable, and the Disregard scheme had relatively better performance.

To show the difficulties of subword-phrase recognition tasks with different labeling schemes, Table 3 displays the F scores of the auxiliary tasks. In general, the number of classes in a sequence labeling problem is related to its difficulty. Thus, the All-Phrase scheme was expected to be the easiest and the Bit-Label scheme the most difficult. As shown in the results in the table, the F scores of the All-Phrase scheme are the highest among these schemes, thereby confirming their easiness in terms of a sequence labeling problem. In contrast, F scores of the other schemes were inferior, but still high enough to aid the generalization performance of the main text classification model.

4.5 Remarks

This section summarizes the findings from our experiment by answering the abovementioned items and introduces the limitations of the proposed method.

(1) The proposed method outperformed the baseline method when the number of classes of a dataset was large and was comparable to them when the number was small. However, datasets with a few classes were also less skewed in the number of instances per class. Therefore, the frequency-based subword-phrase extraction for constructing auxiliary tasks was suitable when datasets had many classes, and the number of instances per class was skewed. A promising outcome is that an auxiliary recognition task in which (pseudo) supervision is generated unsupervised is effective in the MTL-based classification. Therefore, this outcome opens up new possibilities for constructing auxiliary tasks for the MTL-based classification methods on tasks other than text classification.

(2) The proposed method was superior to the SOTA method, RoBERTaGCN, for the R52 and OHS datasets, which contained many classes and where the number of instances per class was skewed. A promising direction to overcome the inferiority of the proposed method in the other datasets is to utilize RoBERTaGCN as a base model for the proposed method.

(3) The subword-phrase recognition task as an auxiliary task improves text classifications in various datasets. A promising outcome is the usage of phrasal expressions for subwords, which which needs more attention in the literature.

(4) To handle common subword-phrases among document classes, the Bit-Label scheme, which encodes dependence of subword-phrases in a bit sequence that can represent all combinations of appearing classes, and the Disregard scheme, which ignores common subword-phrases, were the best. The higher the number of classes (e.g., R52), the better the classification performance using the Bit-Label scheme. Contrastingly, the smaller the number of classes (e.g., R8 and OHS), the better the Disregard scheme performance.

Consequently, when the number of classes is large, and the number of instances for document classes is skewed, the MTL-based text classification suffers from the class imbalance problem, which is still an open problem in the general text classification tasks domain. This domain showcases some promising results by using subword-phrase recognition tasks, whose labels are obtained in an unsupervised manner. However, at the same time, the classification performance still leaves a lot to be desired. Therefore, future studies should seek more effective auxiliary tasks to deal with the class imbalance problem.

5 Conclusion

We proposed an MTL-based text classification framework using auxiliary tasks with lower human and financial costs by creating auxiliary task labels unsupervised. We also sought to ascertain the possibility of phrasal expressions of subwords called subword-phrases to utilize subword-based neural language pre-trained models. As an implementation of our framework, we extracted subword-phrases in terms of their frequency of occurrence and labeled them into documents in three different ways. Our experimental evaluation for text classification using five popular datasets highlighted the effectiveness of the subword-phrase recognition as an auxiliary task. It also showed comparative results with RoBERTaGCN which is the state-of-the-art method.

The main conclusions of this paper are: an auxiliary recognition task in which pseudo supervision is generated in an unsupervised manner is effective in MTL-based classification, and opens up the possibility of constructing auxiliary tasks for MTL-based classification methods for classification tasks other than text classification, and phrasal expressions for subwords (subword-phrase) can be helpful in text classification.

Acknowledgment

This work was partly supported by the Grants-in-Aid for Academic Promotion, Graduate School of Culture and Information Science, Doshisha University, JSPS KAKENHI Grant Number 19H01138, 19H04218, and 21H03555, and JST, the establishment of university fellowships towards the creation of science and technology innovation, Grant Number JPMJFS2145.

References

[1] C. Apté, F. Damerau, and S. M. Weiss. Automated Learning of Decision Rules for Text Categorization. *ACM Transactions on Information Systems*, 12(3):233–251, 1994.

[2] A. Benayas, R. Hashempour, D. Rumble, S. Jameel, and R. C. De Amorim. Unified Transformer Multi-Task Learning for Intent Classification With Entity Recognition. *IEEE Access*, 9:147306–147314, 2021.

[3] Q. Bi, J. Li, L. Shang, X. Jiang, Q. Liu, and H. Yang. MTRec: Multi-Task Learning over BERT for News Recommendation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2663–2669, May 2022.

[4] J. Blitzer, M. Dredze, and F. Pereira. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, 2007.

[5] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, 2020.

[6] R. Caruana. Multitask Learning. *Machine Learning*, 28(1):41–75, 1997.

[7] O. de Gibert, N. Pérez, A. G. Pablos, and M. Cuadros. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, 2018.

[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 2 (Industry Papers)*, pages 4171–4186, 2019.

[9] S. Graham, Q. D. Vu, M. Jahanifar, S. Raza, F. A. Afsar, D. R. J. Snead, and N. M. Rajpoot. One model is all you need: Multi-task learning enables simultaneous histology image segmentation and classification. *Medical Image Analysis*, 83:102685, 2023.

[10] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, 2015.

[11] C. Li, J. Hu, T. Li, S. Du, and F. Teng. An effective multi-task learning model for end-to-end emotion-cause pair extraction. *Applied Intelligence*, 53(3):3519–3529, 2023.

[12] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, and L. He. A Survey on Text Classification: From Traditional to Deep Learning. *ACM Transactions on Intelligent Systems and Technology*, 13(2):31:1–31:41, 2022.

[13] X. Li and D. Roth. Learning Question Classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002.

[14] Y. Lin, Y. Meng, X. Sun, Q. Han, K. Kuang, J. Li, and F. Wu. BertGCN: Transductive Text Classification by Combining GNN and BERT. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1456–1462, Online, Aug. 2021.

[15] M. Lippi, P. Palka, G. Contissa, F. Lagioia, H. Micklitz, G. Sartor, and P. Torroni. CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. *Artifcial Intelligence and Law*, 27(2):117–139, 2019.

[16] P. Liu, X. Qiu, and X. Huang. Deep Multi-Task Learning with Shared Memory for Text Classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 118–127, 2016.

[17] P. Liu, X. Qiu, and X. Huang. Recurrent Neural Network for Text Classification with Multi-Task Learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2873–2879, 2016.

[18] X. Liu, J. Gao, X. He, L. Deng, K. Duh, and Y. Wang. Representation Learning Using Multi-Task Deep Neural Networks for Semantic Classification and Information Retrieval. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 912–921, 2015.

[19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692, 2019.

[20] I. Loshchilov and F. Hutter. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[21] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150. Association for Computational Linguistics, 2011.

[22] Y. Mao, Z. Wang, W. Liu, X. Lin, and P. Xie. MetaWeighting: Learning to Weight Tasks in Multi-Task Learning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3436–3448, 2022.

[23] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao. Deep Learning-based Text Classification: A Comprehensive Review. *ACM Computing Surveys*, 54(3):62:1–62:40, 2021.

[24] R. Qi, M. Yang, Y. Jian, Z. Li, and H. Chen. A Local context focus learning model for joint multi-task using syntactic dependency relative distance. *Applied Intelligence*, 53(4):4145–4161, 2023.

[25] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-Shot Text-to-Image Generation. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8821–8831, 2021.

[26] L. Ramshaw and M. Marcus. Text Chunking using Transformation-Based Learning. In *Third Workshop on Very Large Corpora*, 1995.

[27] F. Sebastiani. Machine Learning in Automated Text Categorization. *ACM computing surveys*, 34(1):1–47, mar 2002.

[28] O. Sener and V. Koltun. Multi-Task Learning as Multi-Objective Optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 525–536, 2018.

[29] R. Sennrich, B. Haddow, and A. Birch. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, Aug. 2016. Association for Computational Linguistics.

[30] T. Tohti, M. Abdurxit, and A. Hamdulla. Medical QA Oriented Multi-Task Learning Model for Question Intent Classification and Named Entity Recognition. *Information*, 13(12):581, 2022.

[31] C. Wu, G. Luo, C. Guo, Y. Ren, A. Zheng, and C. Yang. An attention-based multi-task model for named entity recognition and intent analysis of Chinese online medical questions. *Journal of Biomedical Informatics*, 108:103511, 2020.

[32] M. Xu, K. Huang, and X. Qi. A Regional-Attentive Multi-Task Learning Framework for Breast Ultrasound Image Segmentation and Classification. *IEEE Access*, 11:5377–5392, 2023.

[33] H. Yang, B. Zeng, J. Yang, Y. Song, and R. Xu. A multi-task learning model for Chinese-oriented aspect polarity classification and aspect term extraction. *Neurocomputing*, 419:344–356, 2021.

[34] L. Yao, C. Mao, and Y. Luo. Graph Convolutional Networks for Text Classification. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, pages 7370–7377, 2019.

[35] H. Zhang, L. Xiao, Y. Wang, and Y. Jin. A Generalized Recurrent Neural Architecture for Text Classification with Multi-Task Learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 3385–3391, 2017.

[36] X. Zhang, Q. Zhang, Z. Yan, R. Liu, and Y. Cao. Enhancing Label Correlation Feedback in Multi-Label Text Classification via Multi-Task Learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1190–1200. Association for Computational Linguistics, 2021.

[37] Y. Zhang and Q. Yang. A Survey on Multi-Task Learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2022.

[38] Y. Zhang, N. Zincir-Heywood, and E. Milios. Narrative Text Classification for Automatic Key Phrase Extraction in Web Document Corpora. In *Proceedings of the 7th Annual ACM International Workshop on Web Information and Data Management*, WIDM '05, page 51–58, 2005.

[39] Z. Zhang, W. Yu, M. Yu, Z. Guo, and M. Jiang. A Survey of Multi-task Learning in Natural Language Processing: Regarding Task Relatedness and Training Methods. *CoRR*, abs/2204.03508, 2022.

[40] M. Zhao, J. Yang, and L. Qu. A multi-task learning model with graph convolutional networks for aspect term extraction and polarity classification. *Applied Intelligence*, 53(6):6585–6603, 2023.

Motion Embedded Images: An Approach to Capture Spatial and Temporal Features for Action Recognition

Tri Le[1,3], Nham Huynh-Duc[1,3], Chung Thai Nguyen[1,3] and Minh-Triet Tran[1,2,3]
[1]Faculty of Information Technology, University of Science, VNU-HCM
[2]Software Engineering Lab, University of Science, VNU-HCM
[3]Vietnam National University, Ho Chi Minh City

Keywords: action recognition, motion embedded image, sports dataset, two-stream network

Received: March 22, 2023

The demand for human activity recognition (HAR) from videos has witnessed a significant surge in various real-life applications, including video surveillance, healthcare, elderly care, among others. The explotion of short-form videos on social media platforms has further intensified the interest in this domain. This research endeavors to focus on the problem of HAR in general short videos. In contrast to still images, video clips offer both spatial and temporal information, rendering it challenging to extract complementary information on appearance from still frames and motion between frames. This research makes a two-fold contribution. Firstly, we investigate the use of motion-embedded images in a variant of two-stream Convolutional Neural Network architecture, in which one stream captures motion using combined batches of frames, while another stream employs a normal image classification ConvNet to classify static appearance. Secondly, we create a novel dataset of Southeast Asian Sports short videos that encompasses both videos with and without effects, which is a modern factor that is lacking in all currently available datasets used for benchmarking models. The proposed model is trained and evaluated on two benchmarks: UCF-101 and SEAGS-V1. The results reveal that the proposed model yields competitive performance compared to prior attempts to address the same problem.

Povzetek: Raziskava predstavi model za prepoznavanje človeških aktivnosti iz videov in testira model na novi bazi video posnetkov jugovzhodne Azije.

1 Introduction

The task of human activity recognition (HAR) pertains to the labeling of actions or activities observed within video clips. In recent years, the proliferation of online social platforms has led to an exponential increase in the volume of media data being uploaded, with short-form videos dominating the internet landscape, beginning with Tiktok and now extending to Facebook, Instagram, and Youtube. Consequently, the need for HAR has become increasingly crucial across a range of domains, including content monitoring, classification, and recommendation systems, video retrieval, human-computer interaction, and robotics.

In contrast to a still image, a video clip affords not only static spatial information confined within a single frame but also temporal information that results from integrating spatial information across frames to capture dynamic motions.

There exists a plethora of research investigating the challenging task of video classification. Currently, the majority of high-accuracy results have been obtained using 3D convolutional kernels to capture the temporal information within videos [1][7][3]. Nonetheless, this architecture may be cost-prohibitive to employ in practical scenarios due to its high computational requirements. Consequently, certain approaches prioritize computational efficiency to handle larger datasets, yet may not be suitable for real-world applications [26][15][2]. These methods often necessitate powerful processors to train successfully. Conversely, training Convolutional Neural Networks (ConvNets) to acquire temporal information in videos offers a straightforward, albeit effective alternative. Researchers following this approach vary in their methods for processing original frames, such as fusing temporal information early or late in the network [11], or combining multiple sequential frames to generate optical flow information [18]. Motivated by the positive outcomes of these studies and the effectiveness of ConvNet models in image recognition, we seek to explore the performance of ConvNet models for video classification. Notably, the extraction of temporal information in short videos remains a less explored domain, likely owing to its inherent difficulty. This paper introduces a novel approach for embedding both temporal and spatial features of consecutive video frames into images, thereby enabling effective recognition of the static features of a scene, such as objects, context, and entities, as well as the motion information. Specifically, we incorporate this method into a variant of the two-stream ConvNet model. The first stream leverages the images generated by our approach to detect motion in videos, while the second stream employs a conventional image classification network to recognize spatial information, utilizing single still video frames as inputs. This latter stream aims to identify and preserve any spatial information that might be lacking in the former.

To evaluate the performance of action recognition models, various publicly available datasets such as UCF-101 [19] and UCF Sport [17] have been introduced, containing 101 action and 10 sport classes, respectively. Some datasets attempt to cover a broader range of activities by including more classes[11][12], while others incorporate user-uploaded data from multiple media sources such as Youtube and Vimeo to simulate daily human activities [8][5]. Despite these efforts, most video datasets lack the complexity of videos edited using text, filters, and effects that are prevalent in short-form videos on social networks like Tiktok, Facebook, and Youtube. These limitations can lead to inaccurate benchmarking of models when applied to this new form of video content. In this research, we also aim to collect a novel dataset that includes both non-effected and effected clips. Inspired by previous datasets [17][11], we gathered data within the same Sport category and focused on South-East Asian Game sports. Our dataset, SEAGS_V1, consists of **8** sports classes and **1,168** videos sourced from Youtube and Tiktok. The availability[1] of this dataset will enable researchers to evaluate the performance of their models on a more diverse range of video content.

In this study, we evaluate the performance of our proposed MEI Two-stream network on two widely-used action recognition datasets, UCF-101 and SEAGS_V1. To investigate the potential of our approach further, we also experiment with different backbone architectures and integrate them into an EnsembleNet. Our empirical results demonstrate that our proposed method holds considerable promise in enhancing the accuracy of Activity Recognition on short-form videos.

The content of this paper is organized as follows. In Section 2, we briefly review existing work related to action recognition. Then we present our proposed method in Section 3. We discuss our experiments in Section 4. Finally, the conclusion and future work are discussed in Section 5.

2 Related Work

The early-stage methodologies employed for video classification tasks typically involve a three-stage process. Firstly, visual features of a video segment are extracted densely [20] or at a sparse set of interest points[14]. Secondly, these extracted features are combined into a fixed-sized video-level description. Lastly, a classifier, such as a SVM, is trained on the resulting ”bag of words” representation to discriminate between the pertinent visual classes. Subsequently, ConvNets have replaced all three stages with a single neural network that is end-to-end trainable. However, there are several approaches to augment the connectivity of a ConvNet in the time domain, exploiting local spatio-temporal information[9] [11]. However, these approaches are challenged by the limitations of ConvNets in capturing motion information among frames, leading to the loss of temporal features.

2.1 Two-stream architecture

To mitigate the aforementioned challenge, researchers investigated a novel two-stream ConvNet architecture [18] [21] [25]. This architecture involves feeding the input videos into two distinct streams: the spatial and temporal streams. Each stream employs a deep ConvNet, with softmax scores combined by late fusion. Notably, the inputs for each stream differ slightly. The spatial stream processes individual video frames to recognize actions from still images. In contrast, the temporal stream works on pre-computed optical flow features using optical flow estimation techniques, such as [23].

2.2 Spatial-temporal feature fusion method

The two-stream architecture has inspired numerous studies, with many seeking to improve its performance by focusing on two key areas: the fusion stage and the temporal stream. In an effort to optimize the fusion stage, Feichtenhofer *et. al.* conducted a comprehensive investigation of various approaches to fusing the two networks over space and time [4]. They ultimately discover that fusing a spatial and temporal network at the convolution layer instead of the softmax layer results in comparable performance, while also significantly reducing a substantial number of parameters. Another approach involves using a separate architecture to combine image information. Yue *et. al.* explored two video-classification methods [22] which are both capable of aggregating frame-level ConvNet outputs into video-level predictions: Feature Pooling methods max-pool local information through time, while LSTM's hidden state evolves with each subsequent frame.

2.3 Variations of temporal stream

Various approaches have been explored in an effort to improve the performance of the temporal stream in the two-stream architecture. Zhang *et. al.* investigates the replacement of optical flow with motion vector, which can be obtained directly from compressed videos without additional calculation [24], resulting in a more than 20x speedup compared to traditional two-stream approaches. However, motion vectors tend to lack fine structures and contain noisy and inaccurate motion patterns, leading to a decline in recognition performance. An alternative approach involves learning to predict optical flow using a supervised ConvNet. Ng. *et. al.* proposes a multitask learning model, ActionFlowNet, that trains a single stream network directly from raw pixels to jointly estimate optical flow while recognizing actions with ConvNet, capturing both appearance and motion in a single model [16].

In this study, we build upon the ideas of the two-stream architecture [18] and modify the temporal stream. Rather than relying on optical flow, we introduce a novel approach that embeds motion into the original frames, generating motion-embedded images that retain spatial features in the temporal stream. This is based on the belief that motion and appearance should not be separated. However, the spatial stream is still considered, as our current method for gener-

[1] SEAGS_V1 is currently available online here.

ating motion-embedded images may contain noisy and inaccurate motion patterns caused by background movement.

3 Proposed Method

In this section, we introduce our novel approach called motion embedded image (MEI) and two-stream network. The input video is fed into two distinct streams, the normal and motion streams, as illustrated in Figure 1. The processes in these streams are implemented separately. Prior to being input into the streams, the input can be pre-processed. These inputs are then fed into a ConvNet to perform image classification, and the prediction scores of both streams are fused to produce the final prediction. In the following subsections, we provide comprehensive details of the motion embedding technique, motion stream, normal stream, and fusion stage.

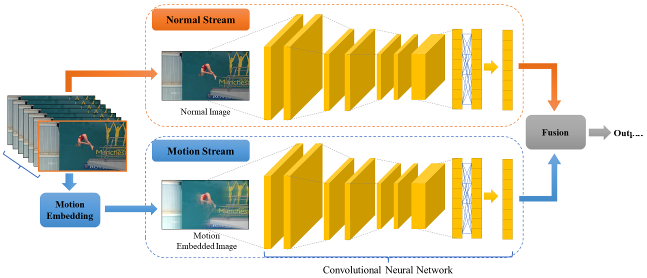


Figure 1: Illustration of our proposed two-stream architecture. Normal stream (top) takes individual frames as inputs, while Motion stream (bottom) requires motion embedded images which are a combination of consecutive video frames. Then, the convolutional neural networks in both streams learn to classify them. Finally, a fusion algorithm is performed to combine normal-motion information. Both streams are end-to-end trainable.

3.1 Motion Embedding

As per the requirements of the Motion stream, the input video frames must undergo a motion embedding stage. Our proposed motion embedding techniques are illustrated in Figure 2, which depict the workflow involved in this stage. The resulting output of this stage is motion-embedded images that convey the direction and order of motion of a single image. Furthermore, we believe that the spatial and temporal information stored simultaneously gives more features for Convolutional Neural Network to learn, which is described in detail in a later sub-section.

All frames extracted from the input video are orderly numbered as T and segmented into batches consisting of N consecutive frames. Each batch is fed into the motion embedding stage, which comprises two components: image processing and combinator. The image processing component is responsible for generating new images from origins, while the combinator aggregates the processed images to create motion-embedded images. It is noteworthy that the aggregation of consecutive frames in a video emphasizes the parts containing static objects and contexts, highlighting the contours of the different stages in the motion that can be easily distinguished from the static parts. The combinator is often dependent on the method used in the image processing component. In the following sub-section, we present our studies about two methods for processing images and their corresponding combinators.

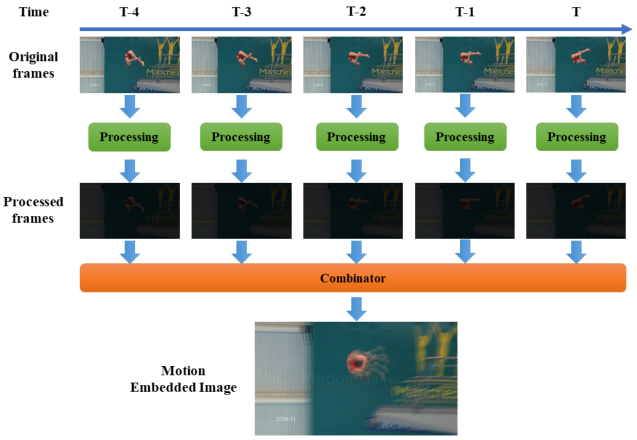


Figure 2: Workflow of our motion embedding technique. The figure illustrates a batch of N=5 consecutive frames from an input video before and after processing which uses the Equal Division method. A combinator, then, merges all processed frames to generate relevant motion embedded images.

3.1.1 Equal division

To ensure that all frames contribute equally to MEI, we divide the values of all pixels in each frame by N. This technique also enables the combinator to keep the pixel values between 0 and 255. The formula for this technique is presented below:

$$processed\_img = original\_img * \frac{1}{N}$$

In the formula, $processed\_img$ and $original\_img$ are 2D arrays representing the pixel values of the processed and original frames, respectively. The operation is performed element-wise.

The combinator we suggest for this method is simply a summation of all processed images. Therefore, the final MEI for a batch concluding at frame T is formulated by the following equation:

$$\text{MEI}_T = \sum_{i=T-N+1}^{T} processed\_img\ i \qquad (1)$$

In Figure 2, a batch of five consecutive frames from an input video is depicted, which is processed through the motion embedding stage using the Equal Division method. As evident from the figure and equations, it is obvious that the final MEI likely presents a stack of images. Due to the identical contributions of all frames to the final image, the motion transitions are presented in a uniform manner throughout the sequence.

3.1.2 Gradient division

The Equal Division method is limited in that it fails to capture the directionality of the motion, as it presents all action steps in an identical manner. To overcome this limitation, we propose the Gradient Division method. This method prioritizes the most recent frame in a batch to serve as the base frame for activity recognition and appropriately weights the contribution of each frame in the batch, with later frames carrying higher weights than earlier ones.

The following describes our proposed formulas for image processing component:

$$sum\_N = \sum_{i=1}^{N} i, \quad contrib = \frac{T \bmod N + 1}{sum\_N}$$

$$processed\_img = original\_img * contrib$$

In the above formula, $processed\_img$, $original\_img$ are 2D arrays of the processed and original frames' pixel values, respectively. The equation is performed element-wisely. The two scalars sum_N, $contrib$ are aimed to calculate the contribution of frame T in a batch of N frames.

The combinator we suggest for this method is similar to the formula 1 for the Equal Division combinator.

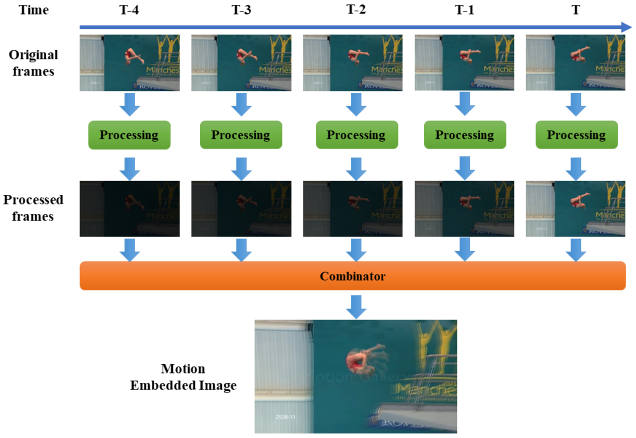


Figure 3: Workflow of our motion embedding technique. The figure illustrates a batch of N=5 consecutive frames from an input video before and after processing which uses the Gradient Division method. A combinator, then, merges all processed frames to generate relevant motion embedded images.

Figure 3 shows a batch of 5 consecutive frames from an input video. It is fed into the motion embedding stage using the Gradient Division method. As shown in the figures and formulas above, the later frames in the batch contribute more to the final output image. This leads to a much better presentation of the direction of action in final motion embedded images. We believe that based on this motion trail, Convolutional Neural Network can learn temporal and spatial information simultaneously.

3.2 Motion stream

The motion stream proceeds in a sequential manner, where batches of N consecutive frames are sequentially fed into the stream. The motion stream operation involves two primary stages. Firstly, the input batch is transformed into an MEI through the motion embedding stage. Subsequently, the generated images are processed by a ConvNet to predict the spatial-temporal features from MEI.

3.3 Normal stream

Initially, we endeavored to investigate the feasibility of employing MEI exclusively for action recognition. However, our experiments revealed that contemporary motion embedding techniques tend to retain motion trails from extraneous objects and backgrounds, resulting in suboptimal outcomes. Consequently, we discerned that static appearance remains a valuable source of information, given its capacity to capture immobile objects without motion trails. Accordingly, we resolved to supplement our approach by adding a normal stream to perform classifications grounded in still images. This stream comprises an image classification ConvNet architecture and can be enhanced by leveraging recent breakthroughs in large-scale image recognition methods [13]. By pre-training this network on a comprehensive image classification dataset, such as the ImageNet challenge dataset, we can further enhance its predictive capabilities.

The normal stream is designed to process individual video frames. In each batch, the most recent frame, referred to as the base frame when using Gradient Division for the motion stream, is extracted and fed into the Convolutional Neural Network (CNN) of this stream.

3.4 Fusion stage

The predictions generated by the two streams of image classification are integrated through a fusion process to produce the ultimate prediction output. At present, our approach to this fusion stage is to compute the arithmetic mean of the predictions, as explicated by formula 2.

$$pred(x) = \frac{normal\_pred(x) + motion\_pred(x)}{2} \qquad (2)$$

where x indicates the input image and $normal\_pred$, $motion\_pred$ and $pred$ present the prediction of normal, motion stream, and the final prediction result, respectively.

4 Experiments and Results

4.1 Dataset

4.1.1 UCF-101

The UCF-101 dataset [19] is a prominent benchmark for evaluating the performance of human action recognition models. The dataset comprises a diverse collection of 101 action classes, spanning over 13,000 clips and 27 hours of video data. Notably, the dataset features realistic user-uploaded videos that capture camera motion and cluttered backgrounds. To evaluate the performance of our approach, we adopt the split-test 01 provided by the authors of this dataset.

4.1.2 SEAGS_V1

We present a novel dataset, SEAGS_V1, that features a diverse mix of effect and non-effect videos.

Our dataset is obtained from a variety of video platforms, including Youtube, TikTok, and Facebook reels. We leverage normal videos as the base data for actions, while short videos with added image effects, text, and stickers serve to enrich the dataset for improved recognition of short effect videos. Figure 6 showcases some examples from our dataset that include text and stickers. Short videos of less than 20 seconds are included in their entirety, except for the intro and outro, while longer videos are manually split into 2-4 segments that are 5-20 seconds in duration.

To facilitate our experiments, SEAGS_V1 is structured in the same manner as UCF-101, with videos organized into folders corresponding to their respective class labels.The name of the video is formatted as

`v_<class label>_<index>.mp4`

We also provide the following files:

`classInd.txt` file contains index of each class label.

`testlist.txt` file contains the path to testing videos accounting for 30% of dataset.

`trainlist.txt` file contains the path to training videos accounting for 70% of dataset.

After data collection, SEAGS_V1 is completed with 8 classes. Each class consists of 100 - 160 videos, each video is between 1 and 20 seconds long. Figures 4, 5 and Table 1 show the statistics of SEAGS_V1 dataset.

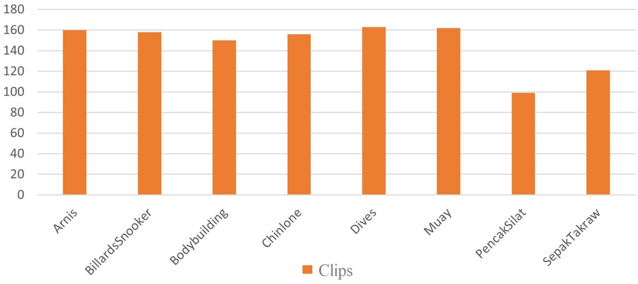


Figure 4: Statistical chart of the clip amount of classes

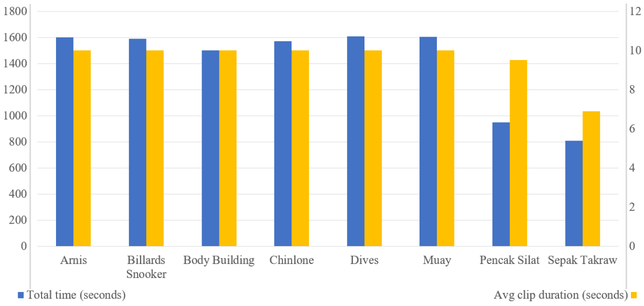


Figure 5: Statistical chart of the total time and average video duration of classes

4.2 Data Augmentation

Upon close examination of our dataset, SEAGS_V1, we figure out that many behaviors are labeled with the same action class, yet differed only in their direction. To further augment the dataset and facilitate learning in these cases, we implemented a data augmentation technique that involves flipping the original images. Figure 6 shows some examples of flipped and original video frames from our dataset.

Table 1: An overview of the SEAGS_V1 dataset

Actions	8
Clips	1169
Total Duration	188 m
Mean Clip Length	9.64 s
Min Clip Length	1.0 s
Max Clip Length	20.0 s
Audio	No

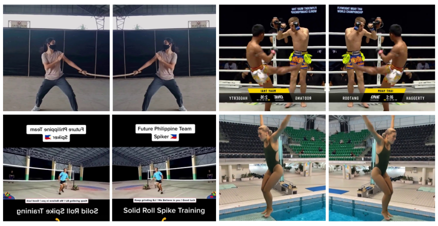


Figure 6: Some flipped and original video frames from dataset SEAGS_V1

4.3 Image classification backbones

For UCF-101, we consider to use EfficientNetB0 as the backbone. For SEAGS_V1, we conduct experiments using a range of backbones, including EfficientNetB0, DenseNet201, InceptionNetV3, ResNet50, and MobileNetV2. Moreover, we explore the potential benefits of ensembling multiple base ConvNet models into a stronger classifier, which we refer to as EnsembleNet, by summing the probability prediction of each model.

$$ensemble\_net(x) = \frac{1}{K}\sum_{k}^{K} base\_net_k(x)$$

where x indicates the input image and K represents the number of base models.

4.4 Motion embedding implementation

We use some specific parameters to create embedded motion images, namely $N = 10$ and interval_frames = 5 for SEAGS_V1 and $N = 10$ and interval_frames = 10 for UCF-101.

Here, interval_frames refers to the distance, in terms of frame count, between two consecutive batches or the distance from the first frame of batch k to the first frame of batch $k + 1$. Each embedded motion image is generated from a batch of N frames. As depicted in Figure 7, a comparison of three types of images - normal image, MEI with

Gradient Division, and Equal Division - highlights the effectiveness of Gradient Division in preserving the direction of motion in activities, whereas Equal Division does not. Accordingly, we employ Gradient Division as the method for the motion embedding process in our experiments. Figure 8 shows some motion-embedded images from both the SEAGS_V1 and UCF-101 datasets.

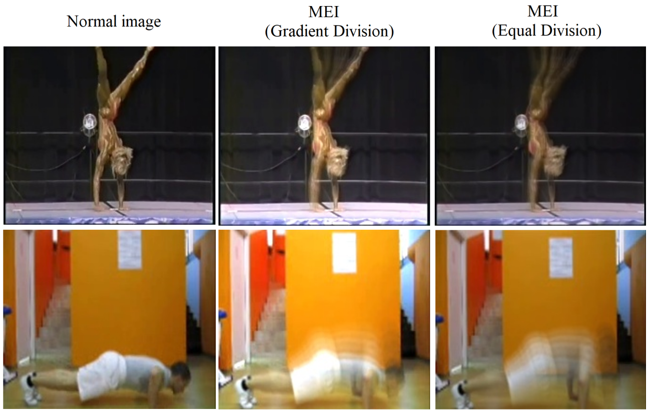


Figure 7: Some examples of normal image (left), MEI with Gradient Division (middle) and with Equal Division (right) from two datasets

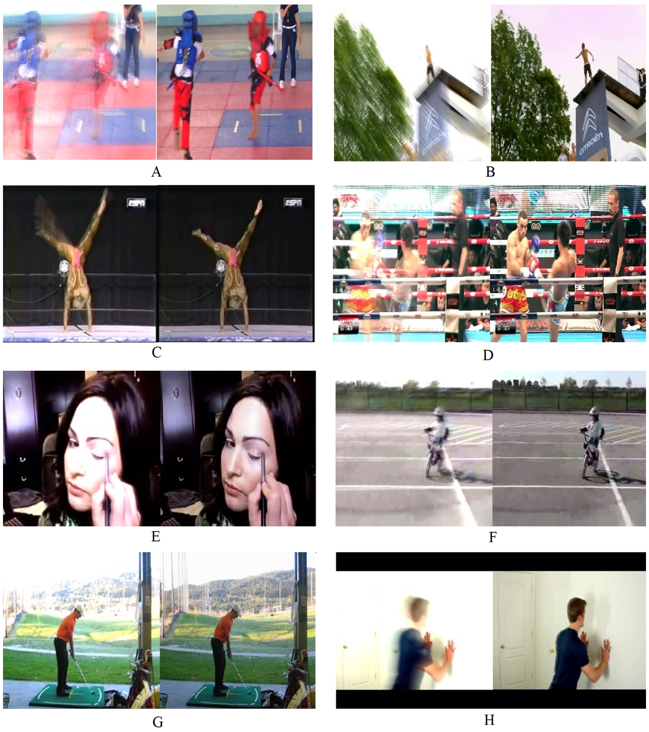


Figure 8: Some motion embedded (left) and its original images (right) from SEAGS_V1 (A, B, C, D) and UCF-101 (E, F, G, H) datasets

4.5 Training

We partition the dataset into training and validation sets at a ratio of 7:3. We conclude the training process once the validation accuracy exceeded 0.9. Notably, training with normal images requires only 10 epochs to achieve the desired validation accuracy, whereas training with MEI takes 50 epochs. Each stream is trained independently, and the probabilities are subsequently fused for prediction purposes.

4.6 Two-stream implementation

We train both the spatial and temporal streams using the same model architecture, albeit independently. The motion stream is fed with the MEIs generated using the parameters specified in the previous section. During testing, the normal stream processes all the last frames of the batches to make predictions.

4.7 Results

Our experimental results on the UCF-101 dataset demonstrate that our proposed method achieved significantly higher accuracy than the initial models developed by Soomro *et. al.* [19], Karpathy *et. al.* [11], and a two-stream model [6]. However, when compared to the original two-stream model [18] and the state-of-the-art approach developed by Wang *et. al.* [10], our method exhibits a noticeable performance gap, as shown in Table 2.

Table 2: Experiment result on UCF-101 dataset (split test 01) (ours with backbone EfficientNetB0)

Model	Accuracy (%)
Soomro et al [19]	43.9
Karpathy et al [11]	65.4
Han et al [6]	68.0
Simonyan et al [18]	88.0
Kalfaoglu et al [10]	**98.69**
Ours (with normal image)	68.54
Ours (with MEI)	67.04
Ours (Two-stream)	70.08

Table 3: Experiment result on SEAGS_V1 dataset with normal image

Backbone	Accuracy (%)
EfficientNetB0	84.9
DenseNet201	89.2
MobileNetV2	87.2
ResNet50	64.1
InceptionV3	86.9
Ensemble (5 base models)	**92.9**

(Done on 1/10 of the total frames of each video)

Overall, the experimental results presented in Tables 2, 3, and 4 suggest that the accuracy of models trained with MEIs is marginally lower than that of models trained with normal images. In particular, the incorrect predictions of MEI-based models are primarily observed in videos with

Table 4: Experiment result on SEAGS_V1 dataset with motion embedded image

Backbone	**Accuracy (%)**
EfficientNetB0	88.3
DenseNet201	87.5
MobileNetV2	81.5
ResNet50	52.7
InceptionV3	85.8
Ensemble (5 base models)	**92.9**

Table 5: Experiment result on SEAGS_V1 dataset with proposed two-stream model

Backbone	**Accuracy (%)**
EfficientNetB0	**90.02**
DenseNet201	89.46
MobileNetV2	88.89
ResNet50	60.11
InceptionV3	88.32

moving contexts, where the MEIs generated from these videos make it difficult for the models to distinguish between actions and context, resulting in suboptimal performance. Figure 8 (B, F) provides examples of poorly generated MEIs from such videos. In contrast, normal images are found to preserve clear visual information among objects, even in the presence of moving contexts.

Conversely, MEIs exhibit a distinct advantage in videos with static or minimally moving contexts, where they can effectively highlight the motion of activities that may not be apparent in normal images. Figure 8 provides examples of such scenarios (A, C, D, H). Hence, the fusion of these two types of images in a two-stream architecture significantly improves the accuracy of the final result on both datasets, as evidenced by the results presented in Tables 2 and 5. Notably, in cases where the motion of activities is relatively consistent, MEIs and normal images exhibit similar characteristics, and the models can effectively learn spatial information. Figure 8 (E, G) provides examples of such cases.

5 Conclusion

In this paper, we propose an approach of applying motion embedded Image (MEI) in a human activity recognition two-stream ConvNet model for short-form videos. We also propose an unprecedented dataset called SEAGS_V1, which consists of both non-effected and effected short videos of 8 local Southeast Asian Sports.

Currently, our experiments on UCF-101 and SEAGS_V1 datasets show that combining the motion stream with the normal spatial stream gives significantly better results than using each stream as an independent model. Moreover, ConvNet models using the ensembled backbone have notably higher accuracy than those using only one backbone. The derived results show a promising potential of the model to advance prediction efficiency in the human activity recognition problem.

Extra training data is beneficial for our model to learn spatial and temporal information, so we are planning to train it on large video datasets such as Sports-1M. Our next direction is to modify the architecture so it can focus more on the activity instead of the whole image and the extracted information will not be diluted. The most important improvement plan is to make the motion stream retain more spatial information so the model only consists of one motion stream and becomes more lightweight.

Acknowledgement

This research is supported by research funding from Honors Program, University of Science, Vietnam National University - Ho Chi Minh City.

References

[1] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. https://doi.org/10.48550/arXiv.1705.07750.

[2] C. Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020. https://doi.org/10.1109/cvpr42600.2020.00028.

[3] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. https://doi.org/10.48550/arXiv.1812.03982.

[4] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016. https://doi.org/10.1109/cvpr.2016.213.

[5] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. https://doi.org/10.1109/iccv.2017.622.

[6] C. Han, C. Wang, E. Mei, J. Redmon, S. K. Divvala, Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue. Yolo-based adaptive window two-stream convolutional neural network for video classification. 2017.

[7] K. Hara, H. Kataoka, and Y. Satoh. Can spatiotemporal 3d cnns retrace the history of 2d

cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. https://doi.org/10.1109/cvpr.2018.00685.

[8] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, 2015. https://doi.org/10.1109/cvpr.2015.7298698.

[9] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013.

[10] M. E. Kalfaoglu, S. Kalkan, and A. A. Alatan. Late temporal modeling in 3d cnn architectures with bert for action recognition. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 731–747. Springer, 2020. https://doi.org/10.1007/978-3-030-68238-$5_4$8.

[11] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. https://doi.org/10.1109/cvpr.2014.223.

[12] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. https://doi.org/10.48550/arXiv.1705.06950.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. https://doi.org/10.1145/3065386.

[14] Laptev and Lindeberg. Space-time interest points. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 432–439 vol.1, 2003. https://doi.org/10.1109/iccv.2003.1238378.

[15] J. Lin, C. Gan, and S. Han. Temporal shift module for efficient video understanding. *CoRR*, abs/1811.08383, 2018. https://doi.org/10.48550/arXiv.1811.08383.

[16] J. Y.-H. Ng, J. Choi, J. Neumann, and L. S. Davis. Actionflownet: Learning motion representation for action recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1616–1624. IEEE, 2018. https://doi.org/10.1109/wacv.2018.00179.

[17] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. https://doi.org/10.1109/cvpr.2008.4587727.

[18] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. https://doi.org/10.48550/arXiv.1406.2199.

[19] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. https://doi.org/10.48550/arXiv.1212.0402.

[20] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR 2011*, pages 3169–3176, 2011. https://ieeexplore.ieee.org/document/5995407.

[21] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao. Towards good practices for very deep two-stream convnets. *CoRR*, abs/1507.02159, 2015. https://doi.org/10.48550/arXiv.1507.02159.

[22] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015. https://doi.org/10.1109/cvpr.2015.7299101.

[23] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l1 optical flow. volume 4713, pages 214–223, 09 2007. https://doi.org/10.1007/978-3-540-74936-$3_2$2.

[24] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang. Real-time action recognition with enhanced motion vector cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2718–2726, 2016. https://doi.org/10.1109/cvpr.2016.297.

[25] Y. Zhao, K. Man, J. Smith, K. Siddique, and S.-U. Guan. Improved two-stream model for human action recognition. *EURASIP Journal on Image and Video Processing*, 2020, 06 2020.

[26] Y. Zhu, Z. Lan, S. Newsam, and A. Hauptmann. Hidden two-stream convolutional networks for action recognition. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pages 363–378. Springer, 2019. https://doi.org/10.48550/arXiv.1704.00389.

Complaints with Target Scope Identification on Social Media

Kazuhiro Ito[1], Taichi Murayama[2], Shuntaro Yada[1], Shoko Wakamiya[1] and Eiji Aramaki[1]
[1]Nara Institute of Science and Technology, Nara, Japan
[2]SANKEN, Osaka University, Osaka, Japan
E-mail: ito.kazuhiro.ih4@is.naist.jp, s-yada@is.naist.jp, wakamiya@is.naist.jp, aramaki@is.naist.jp, taichi@sanken.osaka-u.ac.jp

Keywords: complaint, dataset, Twitter, social media, annotation

Received: March 22, 2023

A complaint is uttered when reality fails to meet one's expectations. Research on complaints, which contributes to our understanding of basic human behavior, has been conducted in the fields of psychology, linguistics, and marketing. Although several approaches have been implemented to the study of complaints, studies have yet focused on a target scope of complaints. Examination of a target scope of complaints is crusial because the functions of complaints, such as evocation of emotion, use of grammar, and intention, are different depending on the target scope. We first tackle the construction and release of a complaint dataset of 6,418 tweets by annotating Japanese texts collected from Twitter with labels of the target scope. Our dataset is available at `https://github.com/sociocom/JaGUCHI`. *We then benchmark the annotated dataset with several machine learning baselines and obtain the best performance of 90.4 F1-score in detecting whether a text was a complaint or not, and a micro-F1 score of 72.2 in identifying the target scope label. Finally, we conducted case studies using our model to demonstrate that identifying a target scope of complaints is useful for sociological analysis.*

Povzetek: Raziskava se osredotoča na analizo pritožb iz 6.418 tvitov z več metodami strojnega učenja.

1 Introduction

[1]A complaint is "*a basic speech act used to express a negative disagreement between reality and expectations for a state, product, organization, or event*" [23, p.195-208]. An analysis of complaints contributes not only to linguistically [30] and psychologically [1, 18] interesting but also beneficial for marketing [17].

Understanding why people are dissatisfied can help improve their well-being by analyzing the situation of their complaints. The methods required to deal with complaints vary greatly depending on whether the target scope of complaints is him/herself, other people, or the environment (e.g., in the workplace, the way of improvement differs when employees are complaining about their own skills or about their work environment). The categorization presented above, regarding the target scope, aligns with James' three psychological categories for the Self as the object of reference [13]: the spiritual Self, the social Self, and the material Self, respectively.

In the field of natural language processing (NLP), there are some studies on how to determine whether a text is a complaint or not [26, 9, 14], or how to identify its severity [15], but no studies have been conducted yet to identify a target scope of complaints, which means the object toward which/whom the complaint is directed. Our study is an attempt to apply a computational approach focusing on a target scope of complaints on social media. More specifically, we emphasize the importance of identifying whether the complaints are intended for the complainer him/herself, for an individual, for a group, or for the surrounding environment.

This paper introduces a novel Japanese complaint dataset collected from Twitter that includes labels indicating the target scope of complaints [2]. We then investigated the validity of our dataset using two classification tasks: a binary classification task (shortly binary task) that identifies whether a text is a complaint or not, and a multiclass classification task (shortly multiclass task) that identifies the target scope of complaints. Furthermore, we apply our target scope classification model to case studies: COVID-19, office work, and the 2011 off the Pacific coast of Tohoku earthquake (we call Tohoku earthquake), aiming to analyze social phenomena.

Our contributions are as follows:

- We constructed a dataset of complaints extracted from Twitter labeled with the target scope of complaints.
- We conducted an experiment with identifying the target scope of complaints and achieved an F1 score of 90.4 in detecting whether a text is a complaint or not, and a micro-F1 score of 72.2 in identifying the target scope label.

[1] This paper is extended version of our study [12] presented in The 11th International Symposium on Information and Communication Technology (SOICT2022)

[2] Our dataset is available at `https://github.com/sociocom/JaGUCHI`

Table 1: Counts and examples of complaint tweets per target scope label in our dataset

Target Scope Label	# of Tweets	Example Tweet
SELF	468	しかしたぶん全部顔とか行動に出ちゃってるから最低なのは自分なんだよね向こうには落ち度はないし勝手に苛ついてるだけだしね (Maybe I'm the one who's the worst because it's all showing on my face and in my actions. It's not the other person's fault, I'm just irritated by myself.)
IND	3,866	わたしが居ないとミルクしまってある場所すらわかんないのかよ (You do not even know where the milk is stored without me?)
GRP	648	価値観の違いかもしれないけど物買うのは3千円でもしぶるのにギャンブルに平気で金突っ込むひとの気持ちがわからない (Maybe it's a difference in values, but I do not understand people who are reluctant to spend even 3,000 yen to buy something, but do not mind excessively spending money on gambling.)
ENV	1,436	保育士の給料上がらないかな〜手取り15〜18じゃやってけないよな (...) 政治家の給料とかより保育士に回してほしいわ、切実に (I wonder if childcare workers' salaries will go up. I can not make it on 15 to 18 take-home pay. (...) I'd really like to see more money spent on childcare workers than on politicians' salaries.)

– We conducted three case studies to demonstrate the usefulness of identifying a target scope of complaints for sociological analysis.

2 Related work

In pragmatics, a complaint is defined as "*a basic speech act used to express a negative disagreement between reality and expectations for a state, product, organization, or event*" [23, p.195‐208]. What makes complaints different from negative sentiment polarity is that complaints tend to include expressions of the breaches of the speaker's expectations [26], and include reasons or explanations [31].

The dataset construction is actively conducted to analyse the substance of complaints. A previous study collected complaints about food products sent to governmental institutions and built an automatic classification model according to the nature of the complaint [9]. The classification classes were set up taking into account the use of customer support, the type of economic activity related, the priority of the treatment, and whether it is under the responsibility of the authority or not. Another study has created complaints dataset with labels for service categories (e.g., foods, cars, electronics, etc.) collected from reply posts to company accounts on Twitter [26]. Another study has also constructed a complaint dataset with four labels [15]: (1) No explicit reproach: there is no explicit mention of the cause and the complaint is not offensive, (2) Disapproval: express explicit negative emotions such as dissatisfaction, annoyance, dislike, and disapproval, (3) Accusation: asserts that someone did something reprehensible, and (4) Blame: assumes the complainee is responsible for the undesirable result. These four categories follow the definitions of the standard in pragmatics [29]. [7] has assigned the intensity of complaints as a continuous value using the best-worst scaling method [20] by crowdsourcing. Another corpus based on the data accumulated by *Fuman Kaitori Center* collects Japanese complaints about products and services [22]. The corpus includes labels about a target of complaints such as product or service names, which is different in granularity from our study.

As mentioned above, although some studies have constructed datasets that collect complaints, they have not yet constructed them that are labeled with a target scope to which complaints are directed.

3 Dataset

3.1 Collection

We constructed a Japanese complaint dataset using Twitter. For our dataset, we collected 64,313 tweets including "# 愚痴 (/gu-chi/)" (a hashtag of a Japanese term for complaints) from March 26, 2006 to September 30, 2021 using the Twitter API[3]. We excluded URLs, duplicates, and retweets, and extracted only those tweets with a relatively low possibility of being a bot. Specifically, we extracted only those tweets for which the posting application was *Twitter for iPad, Twitter for iPhone, Twitter Web App, Twitter Web Client, or Keitai Web*. All hashtags were removed from the text. Tweets with less than 30 characters were excluded. We extracted tweets for each month through a stratified sampling and finally obtained 7,573 tweets, which are

[3] https://developer.twitter.com/

of similar size with datasets recently released for NLP for social media [16, 24, 5, 3, 21].

3.2 Annotation

We annotated the 7,573 tweets with the target scope label. The tweets were divided into three sets (2,524, 2,524, and 2,525 tweets in each set), and three trained external annotators annotated each set.

First stage: Whether the tweet is a complaint or not is identified. Because most of the tweets are complaints owing to the inclusion of "# 愚痴", we remove tweets identified as non-complaints. Following Olshtain's definition [23, p.195‑208], we identified tweets that expressed a negative disagreement between the tweeter's expectations and reality as complaints. Examples of non-complaints tweets removed by this process is shown below.
"If a company is violating the Labor Standards Act, gathering evidence is critical to remedy the situation."
"It's easy to complain, so I'm going to shift my thinking to the positive and creative."
"I came home exhausted again today. But I saw Mt. Fuji for a bit on the train on the way home, and it kind of loosened me up. I thought I was going to cry."

Second stage: We identify the target scope of complaints. We assigned one of four labels, SELF, IND, GRP, and ENV. Although our labels broadly follow James' theory of Self [13], we separate IND (individual) and GRP (group) because we believe that the nature of the complaints differs depending on whether the target is an individual or a group. In the case of individuals, it is associated with abuse, while in the case of groups, it is associated with hate speech. When the target scope was not determined uniquely or was unclear, it was removed from the dataset. We show definitions and examples of labels below.

SELF: A target scope includes the complainer.
e.g., "I have said too much again."

IND: A target scope does not include the complainer, which is one or several other persons.
e.g., "I hate that my boss puts me in charge of his work!"

GRP: A target scope does not include the complainer and has a group.
e.g., "I cannot be interested in people who only think about money."

ENV: A target scope is not human.
e.g., "It's raining today, so I do not feel like doing anything."

As a result of the annotation, among the 7,573 texts, 6,418 were considered as complaints. Among the complaint tweets, the number of labels per target scope is 468 for SELF, 3,866 for IND, 648 for GRP, and 1,436 for ENV. As a result, we collected 6,418 tweets. The agreement ratio (Kappa coefficient) between the annotators and an evaluator was measured to be 0.798 for the binary identification and 0.728 for the four-label classification. Agreement values are between the upper part of the substantial agreement [2]. Figure 1 presents the confusion matrix of human agreement on four classes normalized over the actual values (rows). Examples of text for each target scope label and number of tweets are shown in Table 1.

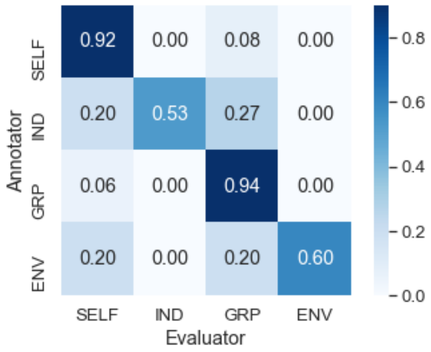


Figure 1: Confusion matrix of annotator agreement on four target scope of complaints.

Table 2: Statistics on the number of characters per label. The label with the highest mean number of characters in the texts is GRP, whereas the label with the lowest mean number of characters in the texts is SELF.

Target Scope Label	Mean	Median	Std
SELF	76.8	74.0	32.2
IND	83.2	83.0	32.4
GRP	87.8	89.0	32.5
ENV	77.8	74.0	33.8
ALL	82.0	81.0	32.8

3.3 Data analysis

We conducted two types of analysis for the contents of the dataset to gain linguistic insight into this task and the data: the number of characters and the emotions. The results of each analysis are shown below.

3.3.1 Number of characters

The average number of characters in the entire dataset is 82.0, and the median is 81.0. The label with the most characters is GRP (mean of 87.8 and median of 89.0), and the label with the fewest characters is SELF (mean of 76.8 and median of 74.0). This suggests that while descriptions of other groups tend to be detailed, those of him/herself have

Table 3: Results of emotion analysis using JIWC. We investigated the average score for each emotion per label. The highest results are in **bold**.

Target Scope Label	Sadness	Anxiety	Anger	Disgust	Trust	Surprise	Joy
SELF	**0.448**	**0.502**	0.774	0.858	**0.591**	0.467	0.459
IND	0.424	0.425	0.846	0.904	0.568	0.457	0.451
GRP	0.407	0.431	**0.861**	**0.954**	0.564	**0.477**	0.444
ENV	0.434	0.490	0.773	0.824	0.545	0.464	**0.482**
ALL	0.426	0.445	0.826	0.888	0.564	0.461	0.458

relatively not in detail. The statistics of the number of characters per label are shown in Table 2. Note that we removed tweets of less than 30 characters in Section 3.1.

3.3.2 Emotion

We examine the relationship between our dataset and emotions, and the differences in emotions between target scope. To do so, we used the Japanese Linguistic Inquiry and Word Count (JIWC) emotion dictionary [4]. This dictionary matches words with seven emotion categories (Joy, Sadness, Anger, Surprise, Trust, Anxiety, and Disgust) based on a translation of Pluchik's emotion wheel [25], obtained from a naturalistic dataset of emotional memories. The scores for each tweet (S_{ij}) were a ratio of the number of emotion terms in each category (W_{ij}), to the total number of terms (tokens; W_i^*) in each tweet:

$$S_{ij} = \frac{W_{ij}}{W_i^*} \log_2(W_{ij} + 1) \quad (1)$$

We used the scores from this emotion dictionary to calculate the emotion score for each tweet in our dataset and investigated the average score for each emotion per label. The results are shown in Table 3.

For SELF, the low value for Anger and high value for Anxiety are consistent with our intuition. When the complainer is him/herself, it can be interpreted that Anxiety is stronger than Anger. Disgust is higher for GRP than for IND. This indicates that feelings of Disgust are stronger for groups than individuals. In the case of Anger, both IND and GRP are high.

3.3.3 Topic

To investigate whether it is possible to extract the detailed contents of complaints in our dataset, we analyzed tweets' topics using the Latent Dirichlet Allocation (LDA), a kind of topic model [4]. The number of topics is set to 8, and LDA is applied only to nouns with two or more Japanese characters. Table 4 shows each topic and assigned words.

The following is an interpretation of the topics. Some of the topics are work-related (Topics 1, 3, 4, and 5), suggesting that work is the majority of complaints posted on Twitter. Among work-related topics, there were topics related to mental health (Topic 3), including "mood," "stress," and "hospital," and topics related to family (Topic 1), including "husband" and "children," which were divided into several tendencies. The other topic focused on COVID-19 (Topic 8), which includes "COVID-19" and "mask." Although only recent tweets are relevant to this topic, it is suggested that many such complaints had been posted intensively.

4 Experiment

4.1 Settings

In this section, we demonstrate the validity of the dataset using two types of classification tasks: a binary task (2-way) that identifies whether a text is a complaint and a multiclass task (4-way) that classifies the target scope of complaints. These tasks correspond to the first and second stages of annotation, respectively.

We employ two types of machine learning models: Long Short-Term Memory (LSTM) [11] and Bidirectional Encoder Representations from Transformers (BERT) [6]. The BERT model is a fine-tuned version of a model pretrained on the Japanese version of Wikipedia published by Tohoku University[5].

Before training, the dataset was preprocessed into lowercase, and all numbers were replaced with zeros. We split the dataset, into training, validation, and test sets (7:1.5:1.5). When we split the dataset the label distribution was maintained.

We set each parameter of the LSTM model as follows: the number of dimensions of the word embedding representation is 10, the number of dimensions of the hidden layer is 128, cross-entropy is used as the loss function, a Stochastic Gradient Descent (SGD) was applied as the optimization method, the learning rate is 0.01, and 100 epochs are used. We also set each parameter of the BERT model as follows: The maximum number of tokens per tweet is 128, the number of batches is 32, Adam is used as the optimization method, the learning rate is 1.0×10^{-5}, and 10 epochs are used. After examination of the validation data, we used the above parameters. Then, for the binary task, we added

[4] https://github.com/sociocom/JIWC-Dictionary

[5] https://github.com/cl-tohoku/bert-japanese

Table 4: The top 5 words per topic (translated from Japanese). Some of the topics are work-related (Topics 1, 3, 4, and 5), suggesting that work is the majority of complaints posted on Twitter. The other topic focused on COVID-19 (Topic 8), which includes “COVID-19” and “mask”.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
husband	child	human	company	really	why	without saying	angry
child	movie	workplace	husband	stupid	friend	adult	COVID-19
boss	parents’ house	mood	world	vacation	everyday	money	cry
mood	block	stress	place	word	child	senior member	forbidden word
senior member	article	hospital	mother	company	meal	staff	mask

6,000 tweets to the dataset that were randomly sampled and removed complaints according to our annotation method.

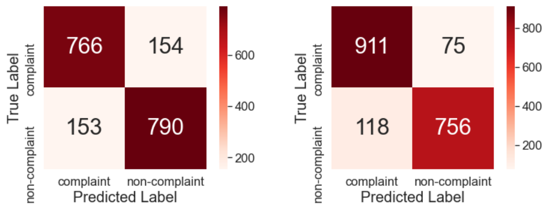
4.2 Metrics

We report predictive performance of the binary task as the mean accuracy, macro-F1 score, and ROC AUC as well as existing complaints study [26]. On the other hand, we report predictive performance of the multiclass task as the micro-F1 score and macro-F1 score.

4.3 Results

4.3.1 Binary task (2-way)

The results of the binary task reach an accuracy level of 83.5, an F1 score of 83.7, and an AUC of 83.5 for the LSTM model, and a level of accuracy of 89.6, an F1 score of 90.4, and an AUC of 89.4 for the BERT model (as shown in Table 5). The confusion matrix of the BERT model has a True Positive rate of 0.92, False Positive rate of 0.14, False Negative rate of 0.08, and True Negative rate of 0.86. For the BERT model, false negatives were reduced in number in comparison to the LSTM model. Figure 2 (a) and (b) show the confusion matrices for the LSTM and BERT models, respectively.



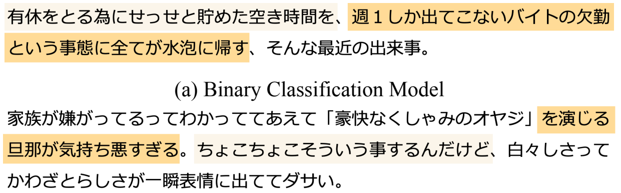
(a) LSTM model (b) BERT model

Figure 2: Confusion matrices of the binary task (2-way).

Table 5: Results of the binary and multiclass tasks. The BERT model outperformed Major Class and the LSTM model for each metric. The bold font indicates the best score for each evaluation metric.

Task	Metric	Major Class	LSTM	BERT
Binary	Accuracy	51.7	83.5	**89.6**
	F1 score	69.3	83.7	**90.4**
	AUC	50.0	83.5	**89.4**
Multiclass	micro-F1 score	62.1	51.7	**72.2**
	macro-F1 score	19.2	30.1	**54.5**

We are interested in what types of tokens our complaint model tries to capture. To interpret the behavior of the model, we used LIME [28], a method for explaining machine learning models, to create a visualization. We visualize the attention weights extracted from BERT model for the following example (translated from Japanese): “Recently, I had an encounter where all the free time I worked hard to make for a paid vacation was wasted because of the absence of a part-time worker who comes to work only once a week.” We observed that the model paid attention to the expression “wasted because of the absence of a part-time worker who comes to work only once a week” for classification (as shown in Figure 3). In this example, the reason was the cause of the complaint, suggesting that our model pays attention to the same part as human intuition.


有休をとる為にせっせと貯めた空き時間を、週１しか出てこないバイトの欠勤という事態に全てが水泡に帰す、そんな最近の出来事。

(a) Binary Classification Model

家族が嫌がってるってわかっててあえて「豪快なくしゃみのオヤジ」を演じる旦那が気持ち悪すぎる。ちょこちょこそういう事するんだけど、白々しさってかわざとらしさが一瞬表情に出ててダサい。

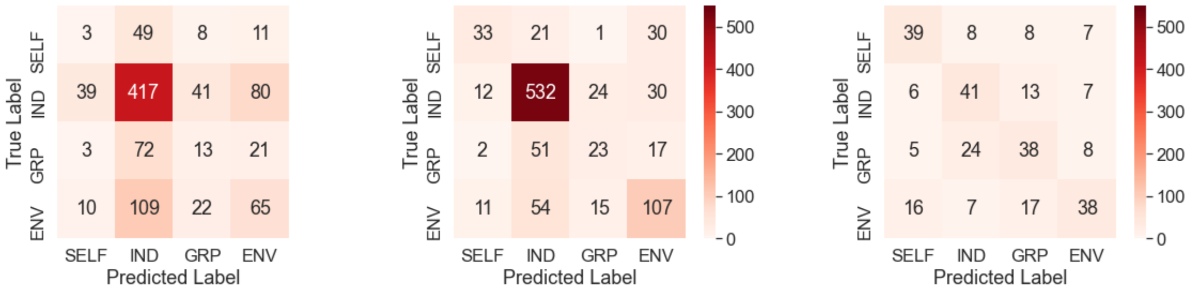
(b) Multi Classification Model

Figure 3: Visualization of the attention weights for the sample sentences in our binary (a) and multi (b) classification models. The orange line highlights the cue of classification. For (a), highlighted words are “wasted because of the absence of a part-time worker who comes to work only once a week.” For (b), highlighted words are “The husband who plays the role of ... too disgusting.”

4.3.2 Multiclass task (4-way)

The results of the multiclass classification task are a micro-F1 score of 51.7 for the LSTM model, and a micro-F1 score of 72.2 for the BERT model. Figure 4 (a) and (b) show the confusion matrices for the LSTM and BERT models, respectively.

In the LSTM model, a relatively large number of tweets are classified as either IND or ENV, reflecting the bias in the number of tweets in the dataset. Although the BERT model mitigates the effect of label bias in the dataset in



(a) LSTM model (b) BERT model (c) BERT with down sampling

Figure 4: Confusion matrices of the multiclass task (4-way). The LSTM model classified a relatively large number of tweets as IND or ENV. The results likely reflect the bias in the number of tweets in the dataset. The BERT model mitigates the effect of label bias in the dataset in comparison to the LSTM model. The BERT model with down sampling results show little bias among the labels.

Table 6: Examples of error cases in the binary task.

ID	Complaint Label		Tweet
	True	Predicted	
(1)	non-complaint	complaint	お仕事終わり!定時で上がれたけど、フィットネスに行くかヤフオクの発送か...。明日は遅番だからジム行くのが得策。来週まで行けないし。(I finished the work! I was able to leave work on time, but I don't know if I should go to the fitness center or ship the Yahoo Auction... I have a late shift tomorrow, so going to the gym is in my best interest. I can't go to there until next week.)
(2)	non-complaint	complaint	何か作りたいなーという気分が出て来ただけマシかなーと思う昨今。風邪の熱に浮かされてるだけかもしれないが。フォトショ起動するのもめんどくさいモードだけど。うん。(I think it's better that I feel like making something these days. I may just be suffering from a fever from a cold. Although I'm too lazy to start up Photoshop right now.)
(3)	non-complaint	complaint	今日は寝坊して大変だったから早め（でももう 0 時;）に寝よう。お休みなさい! (I overslept and had a hard time today, so I'll go to bed early (but it's already midnight;). Good night!)
(4)	complaint	non-complaint	今、カラオケに行ってるらしい。職場にコロナ持ち込まないでねー!! 感染者出たら、あなたの責任ですから! (Now they are going to karaoke, I heard. Don't bring coronavirus into the workplace! If anyone gets infected, it's your fault!)
(5)	complaint	non-complaint	感情豊かですねって、その状況、人に合わせて自分を作ってんだよ (People tell me I'm very emotional, but I make myself fit the situation and the people around me.)

comparison to the LSTM model, the accuracy per label shows that SELF tend to be misclassified as ENV. This reflects the fact that it is difficult to classify SELF and ENV because they have the common tendency to omit the target scope in statements about themselves. The accuracy of GRP is relatively low because when a complainer refers to a group that does not include him/herself, the complainer does not always use words that explicitly express that targets are multiple. In short, the LSTM model greatly outperformed the major class results in macro-F1, and the BERT model somewhat mitigated the bias in the number of labels that affected the LSTM classification results, further improving the macro-F1.

As well as binary task, we show the visualization of what types of tokens our complaint model tries to capture for the following example (translated from Japanese): "The husband who plays the role of "a man sneezing boldly" even though he knows his family doesn't like it is too disgusting. He does it occasionally, and it's so dull because it's so artificial and it shows on his face". This tweet was identified

Table 7: Examples of error cases in the multiclass classification task.

ID	Target Scope Label		Tweet
	True	Predicted	
(6)	SELF	IND	あー、でも休みの日とか、歩いてる時とか、ショッピングの時にアイディア浮かぶかも。もう、おっちゃんアイディア出ないから、もっと若い人に頑張って欲しいなぁ。(Maybe ideas happen when I’m on vacation, or walking, or shopping. As an old man, I can’t come up with any more ideas so I wish more young people would try their best.)
(7)	SELF	ENV	頑張っても報われないし人間関係でいつもとん挫するしどうすりゃいいのかわかんないな、もう (I don’t know what to do because my hard work is not rewarded and I always fail in personal relationships.)
(8)	GRP	IND	とある it 企業のデバッガーとして勤めてますが、今日だけは言わせてください。デバッガーを馬鹿にするな。(I work as a debugger for an IT company, and let me say this today. Don’t mock debuggers.)
(9)	ENV	GRP	ニキビ死ねーーーーーーーっっっ!!!!!!!! お前のせいでブスさ倍増すんだよクソ野郎!!!!!!!! (Pimples go away!!!!!!!!!!!!!!!! You make me look twice as ugly, damn you !!!!!!!!)

as IND by our model. The model paid the most attention to the words “The husband who plays the role of ... too disgusting” for classification (as shown in Figure 3). These words clearly illustrate the target of the complaint, “husband”, and the feeling of “too disgusting” for that person, thus the cues to which the model assigned the labels are clearly interpretable to us.

4.4 Downsampling

Because the error in our multiclass task might be highly influenced by the unbalanced labels of the dataset, we experimented with a dataset with down sampling. We negatively sampled the number of data for labels other than SELF to approximately equal the number of labels for SELF, which has the fewest number of labels. For this experiment, we employ the BERT model and the settings are equal to Section 4.2. The result is a micro-F1 score of 55.3 and a macro-F1 score of 55.5. The results, as illustrated in Figure 4(c), indicate little bias among the labels. This result still shows a relatively high level of confusion between IND and GRP, suggesting that these pairs of labels tend to be similar languages. In addition, there were relatively many cases where ENV tweets were classified as SELF, suggesting that this error may be due to the omission of the target to which the complaint is directed (See Section 4.5).

4.5 Error analysis

4.5.1 Binary task (2-way)

Although the BERT model showed a high score of F1 score of 90.4, the model could not classify tweets correctly in some cases. The examples of error cases are shown in Table 6.

(1), (2), and (3) in Table 6 show the results of False Positive. In the example of (1), although the tweeter writes an expression that is not sure about the choice, it is labeled as NEGATIVE in the true data because It does not contain any negative emotions related to the complaint. In the example of (2), although the word “lazy”, which is closely related to complaints, appear in the sentence, the expression “I think it’s better” is the intent of the entire sentence. In the example of (3), the word “overslept” indicates an unfavorable situation, but the whole sentence is not a complaint because it is simply a tweet indicating the intention to go to bed early. In all of these cases, although negative elements are used in some parts of the tweets, the purpose of the tweet is other than just complaining. These tend to be False Positive.

On the other hand, in the case of (4) and (5) in Table 6, the results are False Negative. The example of (4), syntactically, it is a tweet indicating a kind of request to the target scope, but semantically it is a sentence accusing the target of going out to play. The tweet in (5), tweeter corrects an error in the target’s perception and intends to express that he/she is feeling uncomfortable. As in these examples, there are often cases in which there is no explicitly complaint language or syntax in the tweets, but words appear that semantically imply a complaint.

4.5.2 Multiclass task (4-way)

We use the results of the BERT model with high accuracy to analyze error cases. The examples of error cases are shown in Table 7.

In many cases, the model predicts tweets as IND or ENV whose true labels are SELF. For example, in (6) in Table 7, there are two possible error factors: first, if the model focused on the sentence “I want more young people would try their best” and recognized “young people” as the target, it would be a false identification because the tweeter him/herself is the target scope for the purpose of the tweet. The second is that the tweeter, who is the true target scope, is paraphrased as "old man," and thus this word is perceived as if he were a third party. Example (7) is a tweet that targets him/herself, which the model predicts as a label for ENV, since the scope of the tweet is not explicitly stated. Also, the model predicts tweets as IND or ENV whose true labels are GRP. In example of (8), although it can be inferred from the context that there is more than one person who is the target scope of the complaint, it is difficult to determine from the text whether the number is singular or plural, because there is no noun specified that indicates the target scope of the complaint. In example of (9), the use of the expression "go away" for a non-living target, commonly used to call out to a human, results in the incorrect identification of the target as a human being. Overall, the model tended to misclassify tweets that implied the target scope, which could only be inferred from extra-textual knowledge or the tone of the comments.

5 Case studies

We apply the constructed classification model of a target scope of complaints to tweets related to COVID-19, office work, and Tohoku earthquake to show that it is useful for sociological analysis.

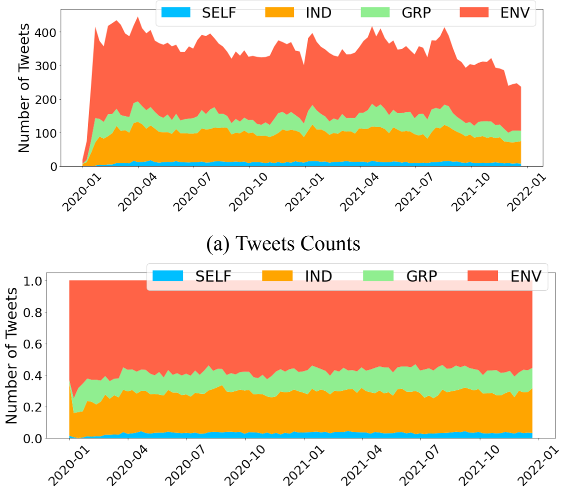
5.1 Case 1: COVID-19

We obtained 698,950 Japanese tweets including "コロナ (/ko-ro-na/)" which is a Japanese word for COVID-19 from January 1, 2020 to December 31, 2021 using the Twitter API.

The time series data presented in Figure 5 show that ENV accounted for a large ratio of cases during the early stages of the pandemic, and that this ratio decreased over time. In the tweets classified as IND or GRP, there were many complaints for others whose views on COVID-19 were different from those of the complainer, whereas in the tweets classified as ENV, there were many complaints for SARS-COV-2 and life during the pandemic. The examples of tweets labeled as each label is shown in Table 8.

In addition, To confirm our hypothesis that a content of complaints varies depending on a target scope, we analyzed the topics of the tweets using the Latent Dirichlet Allocation (LDA), a kind of topic model [4]. The number of topics is set to 16, and LDA is applied only to nouns and adjectives. Table 9 shows the five characteristic topics and five words extracted from the top 10 words per topic. The words that appear in topics about tweets labeled SELF include a number of adjectives such as "afraid," "happy," and "sad," expressing their state of mind. IND is closely related to the tweeter's personal relations, such as "girlfriend," "family," and "parents' house." Complaints about GRP tend to target public things, such as "government," "politics," "Olympics," and "celebrity." ENV frequently contains words related to the services of their customers, such as "lesson," "movie," "vaccine," and "news."

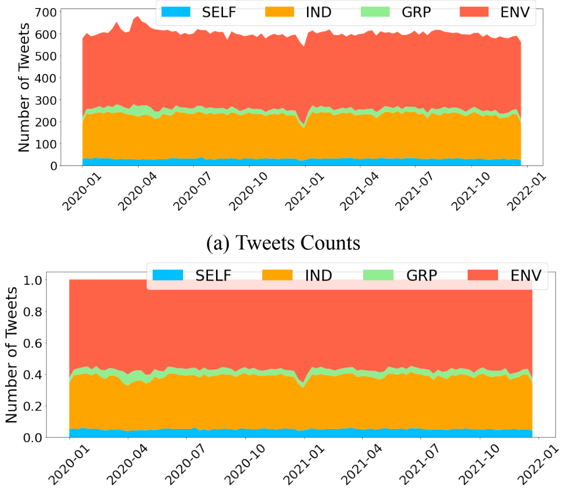
The differences in topics per label showed a certain interpretability, suggesting that automatic classification of a target scope of complaints at the granularity of our dataset also contributes to a categorization of the content of complaints.



(a) Tweets Counts

(b) The ratio of tweets labeled with each label

Figure 5: Time series data on the number of tweets per target scope of complaints related to COVID-19. ENV accounted for a large proportion of cases during the early stages of the pandemic, and this proportion decreased over time.



(a) Tweets Counts

(b) The ratio of tweets labeled with each label

Figure 6: Time series data on the number of tweets per target scope of complaints related to office work. There were few changes in the number of complaints per target scope over time.

Table 8: The examples of tweets related to COVID-19 labeled as each label

Target Scope Label	Tweet
IND	旦那ね、色んなところで営業回ってる人だからよく風邪ひいたり熱出たりすんの。手洗いうがいしてねって言ってもしねぇの。こいつのことこれからコロナさんって呼ぶことにした。(My husband is a salesman who goes around to various places so he often catches a cold or gets a fever. I tell him to wash his hands and gargle, but he doesn't. I've decided to call him Mr. COVID from now on.)
	一生、平行線なんでもういいんじゃないですか。あなたは、コロナは大したことないと思ってる、私は違う。これでいいですよ。(All along, it's failed to reach an agreement, so I think we're done. You think COVID-19 is no big deal, I don't. I'm fine with this.)
ENV	コロナが長引くと永遠に子供に会えなくなります子供はその環境に馴染んでしまうからうちは何とか line で繋げようとしてるけど、もう手遅れなんでそれは悲しいこと (If the situation with COVID-19 is prolonged, we won't be able to see our child forever ... We are trying to connect with them via LINE so that they don't get used to that environment, but it's too late now, and that's sad ... .)
	ホント疲れちゃったし、我慢してることも多いから辛いよコロナ禍じゃなきゃとっくに東京とかも行ってるし、何よりライブ出来てただろうしね (It's hard because I'm really tired and I have to endure so much ... . If it wasn't the situation with COVID-19, I would have been in Tokyo by now, and more importantly, I would have been able to go to live shows.)

Table 9: Five characteristic topics and five words extracted from the top 10 words per topic (translated from Japanese). SELF contains many adjectives such as "afraid," "happy," and "sad," expressing their state of mind. IND is closely related to the tweeter's personal relations, such as "girlfriend," "family," and "parents' house." Complaints about GRP tend to target public things, such as "government," "politics," "Olympics," and "celebrity." ENV frequently include words related to the service for which the tweeter is a customer, such as "lesson," "movie," "vaccine," and "news."

Target Scope Label	Words extracted from the top 10 words per topic				
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
SELF	afraid	hobby	natural	meal	a lot
	happy	ruin	stress	really	complex
	painful	symptoms	dislike	word	surprised
	timing	vaccine	tough	patience	result
	sane	wedding	cheerful	sad	life
IND	part-time job	concert	mask	stupid	afraid
	stress	child	parents' house	money	really
	travel	hospital	test	family	you
	disturbed	aftereffect	fool	mother	friend
	promise	girlfriend	afraid	please	bad
GRP	treatment	covering up	Olympics	vaccine	player
	new type	doctor	report	young man	prejudice
	government	politics	afford	governor	train
	success	opinion	slander	criticism	citizen
	demonstration	civil servants	media	celebrity	trash
ENV	lesson	movie	vaccine	news	infection
	cancellation	ticket	afraid	money	pain
	postponement	gym	time	metropolis	universal
	hospitalization	patience	positivity	summer vacation	like
	return to country	really	insurance	dead	closing down

5.2 Case 2: office work

We obtained 731,000 Japanese tweets including a word "仕事 (/shi-go-to/)", which is related to office work from January 1, 2020 to December 31, 2021 using the Twitter API. Note that among the tweets collected in Case 2, 12,626 tweets overlapped with those collected in Case 1.

The time series data presented in Figure 6 show few changes in the ratio of complaints per target scope over time. This suggests that complaints regarding office work tended to be consistent regardless of the social situation. During the year-end and New Year's periods, the overall number of complaints tended to decrease, while the tweets classified as ENV did not decrease during this period.

As in Case 1, we analyzed the topics of the classified tweets in Case 2. Table 10 shows the five characteristic topics and five words extracted from the top 10 words per topic. The same tendency as in Case 1 was observed for all labels except ENV, with higher weights given to adjectives such as "nervous," "anxious," and "sad" for SELF, words indicating personal relations such as "boss," "you," and "husband" for IND, and words indicating public targets such as "idol," "company," and "voice actor" for GRP.

With regard to ENV, while in Case 1, words indicating services to which the tweeter is a customer appeared, in Case 2, words indicating workload or vacation were common, suggesting that the environment in which complaints target varies greatly depending on the domain.

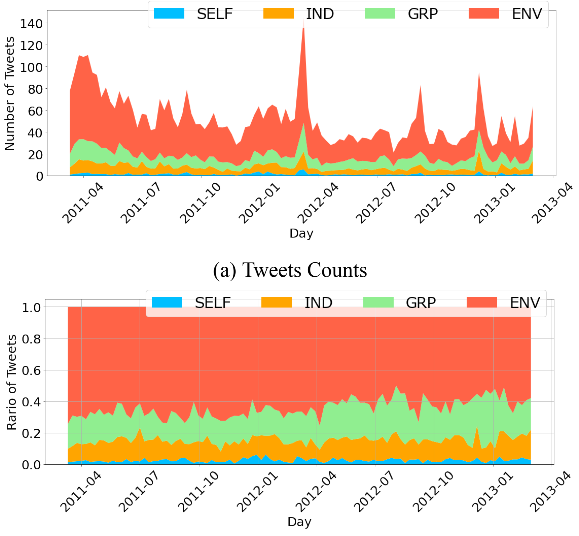
5.3 Case 3: Tohoku earthquake

In Case 1, the time series data show that complaints labeled as ENV accounted for a large proportion of cases during the early stages of the pandemic, but decreased over time, while complaints labeled as IND and GRP are flat over time. This tendency suggests our labels of the target scope of complaints caught phenomenon called "*a paradise built in hell*" [27]. This concept means that victims often exhibit altruistic behavior, engaging in voluntary mutual aid after a disaster. In the case of our classification model, we hypothesize that if the phenomenon of "*a paradise built in hell*" occurs, the ratio of complaints labeled as ENV is high in the early period after the disaster, while the ratio of complaints labeled as IND or GRP increases over time.

We obtained 106,732 Japanese tweets including "東日本大震災 (/hi-ga-shi-ni-ho-n-da-i-shi-n-sa-i/)" which is a Japanese word for Tohoku earthquake from March 11, 2011 to March 10, 2013 using the Twitter API. The time series data presented in Figure 7 show that complaints labeled as ENV accounted for a large ratio of cases during the early period after the disaster and that this ratio decreased over time. In contrast to the complaints labeled as ENV, the ratio of complaints labeled as GRP increased from one year after the disaster. These trends suggest that our classification model for the target scope of complaints can be used to detect the phenomenon of "*a paradise built in hell*" in Tohoku earthquake. The examples of tweets labeled as each label is shown in Table 11.

6 Conclusion & future work

We examined the use of computational linguistics and machine learning methods to analyze the complaints subjects. We introduced the first complaint dataset including labels that indicate a target scope of complaints. We then built BERT-based classification models that achieved F1 score of 90.4 for a binary classification task and micro-F1 score of 72.2 for a multiclass classification task, suggesting the validity of our dataset. Our dataset is available to the research community to foster further research on complaints. While we tried to adjust the unbalanced labels of the dataset by down sampling, it is also possible to adjust it by semi-supervised learning [19, 10] or data augmentation [8]. The validation of methods to improve model performance, including these methods, is our future work.



(a) Tweets Counts

(b) The ratio of tweets labeled with each label

Figure 7: Time series data on the number of tweets per target scope of complaints related to Tohoku earthquake. The complaints labeled as ENV accounted for a large proportion of cases during the early period after the disaster and this proportion decreased over time. In contrast to the complaints labeled as ENV, the ratio of complaints labeled as GRP increased from about one year after the disaster.

Furthermore, from the results of the case studies, we could show the possibility of applying the constructed models to perform sociological analysis. In case study, we applied our model to tweets extracted using queries related to COVID-19, office work, and Tohoku earthquake. In the case of COVID-19, we identified that the ratio of complaints targeting the surrounding environment decreases over time. We found that complaints targeting the surrounding environment and specific individuals were more frequent, with the former being complaints about "others whose views on COVID-19 differ from the tweeter" and the latter being complaints about "the COVID-19 virus and the environment in which infectious disease is spreading." These results suggest most complaints can be divided into two categories: complaints that divide people and complaints generate empathy and cooperation. In the case of the 2011 off The Pacific Coast of Tohoku Earthquake, we showed the potential of our model to detect the phenomenon of "*a paradise built in hell*." These viewpoints show the potential of our dataset as a starting point for sociological analysis.

We also experimented with a topic model for each target scope label as a case study using tweets about COVID-19 and office work, respectively. The distribution of words per topic confirms our hypothesis that the content of complaints varies greatly depending on the target scope. In addition, we observed that the complaints classified by our model as environmentally target scope varied greatly depending on the domain. In the future, as attempted through the case

Table 10: Five characteristic topics and five words extracted from the top 10 words per topic (translated from Japanese). Higher weights were given to adjectives such as "nervous," "anxious," and "sad" for SELF, words indicating personal relations such as "boss," "you," and "husband" for IND, words indicating public targets such as "idol," "company," and "voice actor" for GRP, and words indicating the day of the week, busy season, and vacation for ENV

Target Scope Label	Words extracted from the top 10 words per topic				
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
SELF	nervous overtime hard bothersome painful	human really painful stress get a job	like motivation anxious happiness patience	sad depressed dislike despair adult	lonely busy difficult weekend beautiful
IND	boss every day me information really	you son plan absolutely fool	vacation computer mistake salary husband	bath senior member meal tough friend	meal husband work place bath time
GRP	idol type occupation stupid left-wing	recruitment salary serious company woman	voice actor politics interesting government official knowledge	everybody tough professional on time understanding	doctor crime The Diet last train really
ENV	tired go to work Monday Friday everybody	busy event afraid tough reservation	vacation tough tired study nap	good a fun thing refrain end-of-year dull	game tomorrow weekend happy sleep

Table 11: The examples of tweets related to Tohoku earthquake labeled as each label

Target Scope Label	Tweet
GRP	今、電車に乗っていますが、みんな暑い服着ていますね。だから、余計な電力が必要なのです。もうすぐ東日本大震災から 2 年。もう一度、見つめ直しましょう。あぁあ、電車の空調が入っちゃった。(I'm taking the train now, everyone is wearing hot clothes. So we need extra electric power. It will soon be two years since Tohoku earthquake. Let's look back once again. Ahhh, the air conditioning is on in the train.)
	東日本大震災の被災に関して言えば、未だに復興どころか復旧すら出来ていない所もある。ましてや、福島県の一部県民は、ふるさとへ帰れないままです。選挙をしてる場合でしょうかねぇ。(As for the damage caused by Tohoku earthquake, there are still some areas that have not even been restored, let alone repaired. And some residents of Fukushima Prefecture are still unable to return to their hometowns. I wonder if it's a matter of time to hold elections.)
ENV	勉強横目に東日本大震災のドキュメンタリー見てるけど、恐すぎる。これ今日寝れないやつだ。やっぱ 1 人恐い。。(I'm watching a documentary about Tohoku earthquake while studying, it's too scary. I'm sure I won't be able to sleep today. I'm afraid of being alone..)
	いつ災害がくるかわかりません。東日本大震災のとき、カセットボンベの買い置きがなくて困ったよ。(You never know when a disaster will happen. When Tohoku earthquake happened, I was in trouble because I didn't have any cassette cylinders left over.)

study, we won't only be able to identify a target scope of complaints in a text, but also be able to reveal potential social problems by investigating the temporal change of a target scope of complaints. Furthermore, the analysis results can be applied beyond social media platforms. For example, we are interested in investigating the relationship between workplace well-being and complaints by measuring the number of complaints and their target scope in the daily reports of a particular company. Such applications will be useful for achieving a comfortable life within society.

Acknowledgement

This work was supported by JST-Mirai Program Grant Number JPMJMI21J2, Japan.

References

[1] Mark Alicke et al. "Complaining Behavior in Social Interaction". In: *Personality and Social Psychology*

Bulletin 18 (1992), pp. 286–295. DOI: 10.1177/0146167292183004.

[2] Ron Artstein and Massimo Poesio. "Survey Article: Inter-Coder Agreement for Computational Linguistics". In: *Computational Linguistics* 34.4 (2008), pp. 555–596. DOI: 10.1162/coli.07-034-R2.

[3] Tilman Beck et al. "Investigating label suggestions for opinion mining in German Covid-19 social media". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 1–13. DOI: 10.18653/v1/2021.acl-long.1.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *the Journal of machine Learning research* 3 (2003), pp. 993–1022. DOI: 10.5555/944919.944937.

[5] Yi-Ling Chung et al. "CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 2819–2829. DOI: 10.18653/v1/P19-1271.

[6] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018). DOI: 10.48550/arXiv.1810.04805.

[7] Ming Fang et al. "Analyzing the Intensity of Complaints on Social Media". In: *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 1742–1754. DOI: 10.18653/v1/2022.findings-naacl.132.

[8] Steven Y. Feng et al. "A Survey of Data Augmentation Approaches for NLP". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 968–988. DOI: 10.18653/v1/2021.findings-acl.84.

[9] João Filgueiras et al. "Complaint Analysis and Classification for Economic and Food Safety". In: *Proceedings of the Second Workshop on Economics and Natural Language Processing*. Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 51–60. DOI: 10.18653/v1/D19-5107.

[10] Akash Gautam et al. "Semi-Supervised Iterative Approach for Domain-Specific Complaint Detection in Social Media". In: *Proceedings of the 3rd Workshop on e-Commerce and NLP*. Seattle, WA, USA: Association for Computational Linguistics, July 2020, pp. 46–53. DOI: 10.18653/v1/2020.ecnlp-1.7.

[11] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

[12] Kazuhiro Ito et al. "Identifying A Target Scope of Complaints on Social Media". In: *Proceedings of the 11th International Symposium on Information and Communication Technology*. SoICT '22. Hanoi, Vietnam, 2022, pp. 111–118. DOI: 10.1145/3568562.3568659.

[13] William James. *The Principles of Psychology*. London, England: Dover Publications, 1890.

[14] Mali Jin and Nikolaos Aletras. "Complaint Identification in Social Media with Transformer Networks". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 1765–1771. DOI: 10.18653/v1/2020.coling-main.157.

[15] Mali Jin and Nikolaos Aletras. "Modeling the Severity of Complaints in Social Media". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 2264–2274. DOI: 10.18653/v1/2021.naacl-main.180.

[16] Mali Jin et al. "Automatic Identification and Classification of Bragging in Social Media". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3945–3959. DOI: 10.18653/v1/2022.acl-long.273.

[17] Chul-min Kim et al. "The effect of attitude and perception on consumer complaint intentions". In: *Journal of Consumer Marketing* 20 (2003), pp. 352–371. DOI: 10.1108/07363760310483702.

[18] Robin M. Kowalski. "Complaints and complaining: functions, antecedents, and consequences." In: *Psychological bulletin* 119 2 (1996), pp. 179–96. DOI: 10.1037/0033-2909.119.2.179.

[19] Dong-Hyun Lee. "Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks". In: 2013.

[20] Jordan J Louviere, Terry N Flynn, and Anthony Alfred John Marley. *Best-worst scaling: Theory, methods and applications*. Cambridge University Press, 2015. DOI: 10.1017/cbo9781107337855.

[21] Julia Mendelsohn, Ceren Budak, and David Jurgens. "Modeling Framing in Immigration Discourse on Social Media". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 2219–2263. DOI: `10.18653/v1/2021.naacl-main.179`.

[22] Kensuke Mitsuzawa et al. "FKC Corpus : a Japanese Corpus from New Opinion Survey Service". In: *In proceedings of the Novel Incentives for Collecting Data and Annotation from People: types, implementation, tasking requirements, workflow and results*. Portorož, Slovenia, 2016, pp. 11–18.

[23] Elite Olshtain and Liora Weinbach. "10. Complaints: A study of speech act behavior among native and non-native speakers of Hebrew". In: 1987. DOI: `10.1075/pbcs.5.15ols`.

[24] Silviu Oprea and Walid Magdy. "iSarcasm: A Dataset of Intended Sarcasm". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 1279–1289. DOI: `10.18653/v1/2020.acl-main.118`.

[25] Robert Plutchik. "Chapter 1 - A GENERAL PSYCHOEVOLUTIONARY THEORY OF EMOTION". In: *Theories of Emotion*. Ed. by Robert Plutchik and Henry Kellerman. Academic Press, 1980, pp. 3–33. DOI: `10.1016/B978-0-12-558701-3.50007-7`.

[26] Daniel Preoțiuc-Pietro, Mihaela Gaman, and Nikolaos Aletras. "Automatically Identifying Complaints in Social Media". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 5008–5019. DOI: `10.18653/v1/P19-1495`.

[27] Solnit Rebecca. *A paradise built in hell: The extraordinary communities disaster*. Penguin, 2010.

[28] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 1135–1144. DOI: `10.1145/2939672.2939778`.

[29] Anna Trosborg. *Interlanguage Pragmatics: Requests, Complaints, and Apologies*. De Gruyter Mouton, 2011. DOI: `doi:10.1515/9783110885286`.

[30] Camilla Vásquez. "Complaints online: The case of TripAdvisor". In: *Journal of Pragmatics* 43.6 (2011). Postcolonial pragmatics, pp. 1707–1717. DOI: `10.1016/j.pragma.2010.11.007`.

[31] Guangyu Zhou and Kavita Ganesan. "Linguistic Understanding of Complaints and Praises in User Reviews". In: Jan. 2016, pp. 109–114. DOI: `10.18653/v1/W16-0418`.

Khmer-Vietnamese Neural Machine Translation Improvement Using Data Augmentation Strategies

Thai Nguyen Quoc[1], Huong Le Thanh[1,*] and Hanh Pham Van[2]
[1]School of Information and Communication Technology, Hanoi University of Science and Technology, Vietnam
[2]FPT AI Center
E-mail: thai.nq212642m@sis.hust.edu.vn, huonglt@soict.hust.edu.vn, hanhpv@fsoft.com.vn

Keywords: machine translation, data augmentation, low-resource, Khmer-Vietnamese

Received: March 22, 2023

The development of neural models has greatly improved the performance of machine translation, but these methods require large-scale parallel data, which can be difficult to obtain for low-resource language pairs. To address this issue, this research employs a pre-trained multilingual model and fine-tunes it by using a small bilingual dataset. Additionally, two data-augmentation strategies are proposed to generate new training data: (i) back-translation with the dataset from the source language; (ii) data augmentation via the English pivot language. The proposed approach is applied to the Khmer-Vietnamese machine translation. Experimental results show that our proposed approach outperforms the Google Translator model by 5.3% in terms of BLEU score on a test set of 2,000 Khmer-Vietnamese sentence pairs.

Povzetek: Raziskava uporablja predhodno usposobljen večjezični model in povečanje podatkov. Rezultati presegajo Google Translator za 5,3%.

1 Introduction

Machine translation (MT) is the task of automatically translating text from one language to another. There are three common approaches to MT: rule-based approach [1], statistical-based approach [2, 3], and neural-based one [4, 5, 6]. The rule-based approach depends on translation rules and dictionaries created by human experts. Statistical Machine Translation (SMT) relies on techniques like word alignment and language modeling to optimize the translation process. While SMT can handle a wide range of languages and translation scenarios, it often struggles with capturing complex linguistic phenomena and handling long-range dependencies. With significant advancements in deep learning, Neural Machine Translation (NMT) approaches have shown great potential and have replaced SMT as the primary approach to MT. NMT models capture contextual information, handle word reordering, and generate fluent and natural translations. NMT has gained popularity due to its end-to-end learning, ability to handle complex linguistic phenomena, and improved translation quality. Among all NMT systems, transformer-based MT models [7, 8] have demonstrated superior performance. The key feature of transformer models [8] is their attention mechanism, which allows them to effectively capture dependencies between different words in a sentence. Unlike traditional recurrent neural networks that process words sequentially, transformers can consider the entire input sentence simultaneously. This parallelization significantly speeds up the training process and makes transformers more efficient for long-range dependencies.

One notable limitation of NMT techniques pertains to their reliance on a substantial number of parallel sentence pairs to facilitate model training. Unfortunately, most of language pairs in the world are lack of such a large dataset. Consequently, these language pairs fall under the category of low-resource MT, presenting a challenging scenario for the application of neural-based models in this domain.

Several works have carried out research to solve the low-resource problem in NMT. Chen et al.[9], Kim et al. [10] dealt with the low-resource NMT by using pivot translations, where one or more pivot languages were selected as a bridge between the source and target languages. The source-pivot and the pivot-target should be rich-resource language pairs. Sennric et al. [11], Zhang [12] applied the forward/backward translation approaches to generate parallel sentence pairs by translating the monolingual sentences to the target/source language via a translation system. Then, the pseudo parallel data was mixed with the original parallel data to train an NMT model. A problem in this approach is how to control the quality of the pseudo parallel dataset in order to improve the performance of the low-resource NMT system.

Since NMT requires the capability of both language understanding (e.g., NMT encoder) and generation (e.g., NMT decoder), pre-training language model can be very helpful for NMT, especially low-resource NMT. To do this task, BART model [13] has been proposed to add noises and randomly masked some tokens in the input sentences in the encoder, and learn to reconstruct the original text in the decoder. T5 model [14] randomly masks some tokens and replace the consecutive tokens with a single sentinel token.

To address the low-resource problem in NMT, we propose to fine-tune mBART [15] - a pretrained multilingual Bidirectional and Auto-Regressive Transformers model that has been specifically designed for multilingual applications, including MT. The fine-tuning process is combined with several strategies, including the utilization of back-translation techniques [11] and data augmentation via a pivot language. We propose several data augmentation strategies to augment training data as well as controlling the data quality.

Our proposed approach can be applied to any low-resource language pairs. However, in this research, we evaluate our approach by implementing it with the low-resource Khmer-Vietnamese (Km-Vi) language pair, using a dataset with 142,000 parallel sentence pairs from Nguyen et al. [16]. As far as we know, there is only two works dealing with the Km-Vi machine translation ([17], [18]). Nguyen et al. [17] presented an open-source MT toolkit for low-resource language pairs. However, this approach only used a transformer architecture to train the original dataset, without applying fine-tuning, transfer learning, or additional data augmentation techniques. Pham and Le [18] fine-tuned mBART and applied some data augmentation strategies. In this research, we have extended the work in [18] to improve the performance of the Km-Vi NMT system. The contributions are as follows:

– We propose new methods for data selection based on sentence-level cosine similarity through the bi-encoder model [19] combined with the TF-IDF score.

– We suggest a data generation strategy to generate best candidates for the synthetic parallel dataset.

– To control the quality of augmented data, we propose an “aligned” version to enrich the data and a two-step filtering to eliminate low quality parallel sentence pairs.

The remainder of this paper is organized as follows. Section 2 analyzes various techniques in existing research to address the limitations of low-resource NMT. Section 3 describes our system diagram. Our proposed data augmentation strategies are outlined in Section 4. Section 5 elaborates on the experimental design, whereas Section 6 presents an analysis of the empirical outcomes. Finally, Section 7 concludes our paper.

2 Related work

Pretrain Language Models (PLMs) have proven to be helpful instruments in the context of low-resource NMT. Literature has shown that low-resource NMT models can benefit from the use of a single PLM [20, 21] or a multilingual one [13]. The multilingual PLM is claimed to facilitate more effective learning of the connection between the source and the target representations for translation. These transfer learning methods leverage rich-resource language pairs to train the system, then fine-tune all the parameters on the specific target language pair [22]. The rich-resource language pairs should be in a similar language family to the low-resource ones, to have good results.

Data augmentation is the method of generating additional data, achieved by expanding the original dataset or integrating supplementary data from relevant sources. Various approaches to data augmentation have been explored, including: (i) paraphrasing and sentence simplification [23], (ii) word substitution and deletion [24, 25], (iii) limited and constrained word order permutation [26], (iv) domain adaptation [27], (v) back-translation [11], and (vi) data augmentation via a pivot language [28].

Paraphrasing and sentence simplification [23] offer varied quality, with a risk of introducing semantic changes or losing important information. Word substitution [24] requires careful selection of synonyms to maintain accuracy, while word deletion [25] can introduce noise and requires effective training to handle missing information. Limited and constrained word order permutation [26] suits language pairs with word order variations but requires defining complex constraints based on language characteristics. Domain adaptation [27] addresses the challenge of domain-specific low-resource machine translation, which is not the target of this research. Back-translation [11] has proven successful by generating synthetic source sentences through translating target sentences. However, this approach carries a risk of errors due to imperfections in pre-trained translation models. On the other hand, the pivot-based approach [28] involves translating low-resource language pairs through a high-resource language. This approach relies on good translation quality to and from the pivot language.

Back-translation and pivot-based translation are considered reliable and generalizable approaches when complemented by effective post-processing methods for filtering low-quality data. Therefore, this paper specifically concentrates on utilizing back-translation and pivot-based translation as the selected methods for data augmentation. To improve the quality of the synthetic parallel data generated by these methods, two strategies are employed: (i) data selection and (ii) synthetic data filtering.

Data selection is the process of ranking and selecting a subset from a target monolingual dataset that ensures in-domain as the training data. The objective of this process is to improve the performance of an NMT system for a particular domain. Various techniques for data selection have been proposed in the literature, such as computing sentence scores based on Cross-Entropy Difference (CED) [29, 30], and using representation vectors to rank sentences in the monolingual dataset [31, 32]. Three data selection methods had been implemented by Silva et al. [33], namely CED, TF-IDF, and Feature Decay Algorithms (FDA) [34]. The experimental results pointed out that the TF-IDF method gained the best improvements in both BLEU and TER (Translation Error Rate) scores.

Synthetic data filtering To filter out low-quality sentence pairs, Imankulova et al. [35] proposed a method based on the BLEU measure. This method involves leveraging a source-to-target NMT model to translate the synthetic source sentences into synthetic target sentences. Subsequently, the sentence-level BLEU score is calculated for each sentence pair between the synthetic target sentence and the target sentence, with the ultimate objective of excluding low-score sentences. Koehn et al. [36] proposed another approach based on the sentence-level cosine similarity of two sentences. However, their proposal required an effective acquisition of the linear mapping relationship between the two embedding spaces of the source language and the target one.

Another way to improve translation quality is by using data augmentation methods via a pivot language [28]. This method involves translating sentences from the source language to the pivot one using the source-pivot translation model, followed by translating sentences in the pivot language to sentences in the target language. However, there are certain restrictions associated with this technique. Firstly, the circular translation process increases the decoding time during inference as it can iterate through multiple languages to obtain the desired quality. Secondly, translation errors may arise in each step, which can lead to low-quality translation of the sentence in the target language.

In this paper, we introduce an approach aimed at enhancing the performance of low-resource MT. Our approach incorporates multiple data augmentation strategies alongside various data filtering methods to improve the quality of synthetic data. In the subsequent sections, we introduce these methods in detail.

3 Our system diagram

As previously mentioned, our goal is to propose strategies that can improve the performance of low-resource NMT systems. The proposed approach will be applied for the Km-Vi language pair. To do this, we first fine-tune the mBART50 [37] model with the Km-Vi bilingual dataset.

The mBART model Multilingual BART (mBART) [15] is a sequence-to-sequence denoising auto-encoder that was pre-trained on large-scale monolingual corpora in many languages using the BART objective [13]. The pre-trained task is to reconstruct the original text from the noise one, using two types of noise: random span masking and order permutation. A special variant of mBART called mBART50 [37], has been trained in 50 languages, including Khmer and Vietnamese. Nonetheless, the mBART50's translation quality of the Km-Vi language pairs is low. To deal with this problem, we propose to fine-tune the mBART50 with the Km-Vi bilingual dataset combined with the augmented dataset through several strategies.

Our proposed Khmer-Vietnamese MT system model is described in Figure 1, which incorporates two strategies for data augmentation: (i) back-translation with a dataset in the target language; and (ii) data augmentation via English pivot language. These strategies will be introduced in the next section.

4 Data augmentation strategies

Since the word orders and theirs meaning in machine translation are important, methods such as paraphrasing, simplification, limited and constrained word order permutation cannot provide good parallel sentence pairs.

4.1 Back-translation with a dataset in the target language

Back-translation method proposed by Senrich et al [11] is an useful way to generate additional training data for low-resource NMT. This method leverages an external dataset in the target language, termed the "target-language dataset". It employs a target-to-source NMT model, trained on the original bilingual dataset, to translate this dataset into the source language. The resulting translated sentences are then combined with their corresponding target sentences, creating a synthetic bilingual dataset. However, the dataset's quality generated by this method is not guaranteed. To address this issue, we improve this method by integrating data filtering techniques to the back-translation process. Our proposed method is conducted in three steps as follow:

– **Step 1 - Data selection:** Rank and select sentences from a target-language dataset that is in the same domain as sentences in the original bilingual dataset.

– **Step 2 - Data generation:** Each sentence from the output dataset in Step 1 is translated to k sentential candidates in the source language using the target-to-source NMT model which has been trained by fine-tuning the mBART50 with the original bilingual dataset.

– **Step 3 - Data filtering:** Filter out low-quality bilingual sentence pairs in the synthetic parallel dataset.

We will discuss these three steps in the following sections.

4.1.1 Data selection

For a given dataset D consisting of T sentence pairs in a specific domain, and a set of sentences in a general domain G, the aim of data selection is to rank the sentences in G based on their similarity to the domain of D, then selecting highest-ranked sentences to form a subset that shares the same domain as D. Given that TF-IDF is a popular technique used to identify representative words for a dataset, we can assess whether sentences in G belong to the same domain as D using this measure. In addition to the TF-IDF measure, cosine similarity can be employed to measure the semantic similarity between two sentences based on their

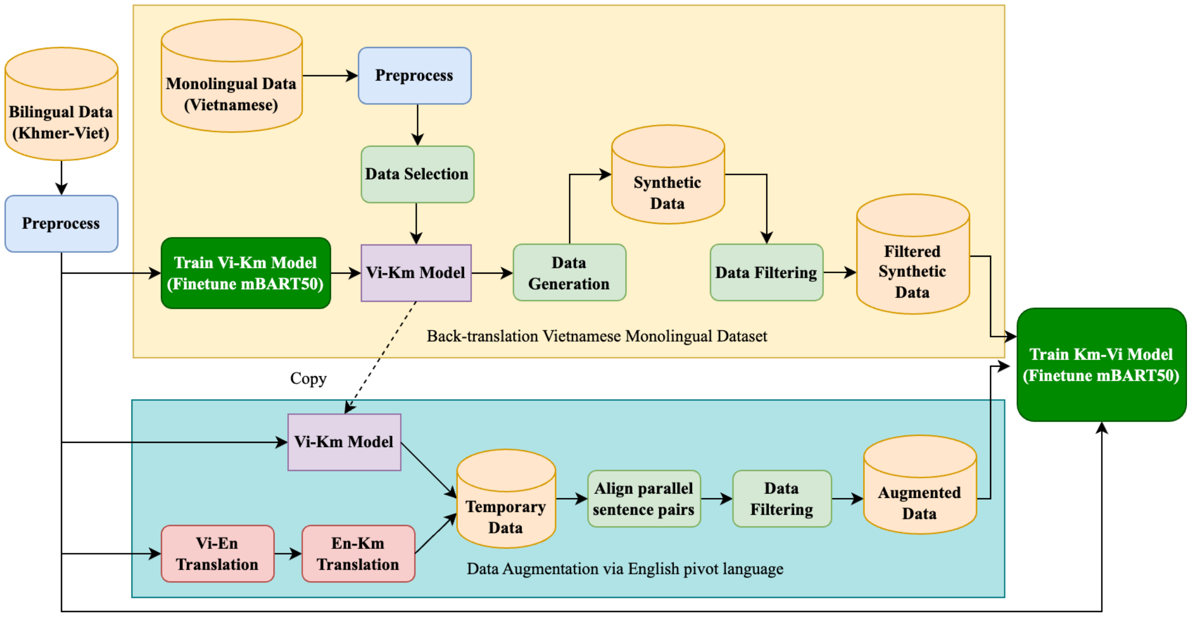

Bilingual Data (Khmer-Viet)
Preprocess
Monolingual Data (Vietnamese)
Preprocess
Data Selection
Train Vi-Km Model (Finetune mBART50)
Vi-Km Model
Data Generation
Synthetic Data
Data Filtering
Filtered Synthetic Data
Back-translation Vietnamese Monolingual Dataset
Copy
Train Km-Vi Model (Finetune mBART50)
Vi-Km Model
Temporary Data
Align parallel sentence pairs
Data Filtering
Augmented Data
Vi-En Translation
En-Km Translation
Data Augmentation via English pivot language

Figure 1: Our proposed Khmer-Vietnamese MT system diagram

semantic vector representations. This enables the identification of sentences in G that share the same domain as the sentences in D. Due to this reason, TF-IDF, cosine similarity, and their combination are utilized for ranking.

Data selection based on TF-IDF score

The term frequency (TF) measures the frequency of a term (word or subword) in a sentence, while inverse document frequency (IDF) is defined as the proportion of documents in the corpus that contain the term. So, TF-IDF score of a word w in a sentence s in G is calculated as:

$$score_w = TF - IDF_w = \frac{F_w^G}{W_s^G} \cdot \frac{T^D}{K_w^D}$$

where F_w^G is the frequency of w in s; W_s^G is the number of words in s; and K_w is the number of sentences in D contain w.

The TF-IDF score of the sentence $s \in G$ is evaluated as:

$$score_s^{(TF-IDF)} = \sum_{i=1}^{W_s^G} score_{w_i}$$

Data selection based on cosine similarity score The cosine similarity score between two sentences is calculated using a Bi-Encoder model [38]. This model includes a PLM combined with a pooler layer to encode each sentence as a sentence-level representation vector. Then, we compute the cosine similarity between these two vectors.

To choose the optimal PLM for the Vietnamese (target) language, we build a test set for the masked language model task, which includes 140,000 Vietnamese sentences from the Km-Vi bilingual dataset. Based on the accuracy of some well-known PLMs (ie, PhoBERT[1], XLM-RoBERTa[2], mDeBERTa[3]) using this dataset (Table 1), XLM-RoBERTa is selected as the PLM for the Bi-Encoder model.

Table 1: Accuracy of some models on the test set for the masked language model task.

Models	Accuracy
PhoBERT	80%
XLM-RoBERTa	87%
mDeBERTa	75%

The cosine similarity score of a sentence s in the G is calculated as:

$$score_s^{(COS)} = \frac{1}{|D|} \sum_{i=1}^{|D|} cos(s, D_i)$$

where $|D|$ is the number of sentences in D; D_i is the i-th sentence in D.

Data selection based on combination score

The combination score is calculated based on the TF-IDF score and the cosine similarity score:

$$score_s = \frac{score_s^{TF-IDF}}{\sum_{j=1}^{|G|} score_{G_j}^{TF-IDF}} + \frac{score_s^{COS}}{\sum_{j=1}^{|G|} score_{G_j}^{COS}}$$

where $|G|$ is the number of sentences in G; G_i is the i-th sentence in G

After assigning these scores to the sentences in the corpus G, the top 120,000 sentences from the target-language dataset with the highest score are selected to translate into the source language based on the target-to-source translation model.

[1] https://huggingface.co/vinai/phobert-base

[2] https://huggingface.co/xlm-roberta-base

[3] https://huggingface.co/microsoft/mdeberta-v3-base

4.1.2 Synthetic data generation

To increase the number of generated sentence pairs, each sentence from the target-language dataset is translated into k candidate sentences in the source language using the beam search (k is beam size) or top-k sampling method. As a result, k bilingual sentence pairs are created for each sentence in the target-language dataset. At this step, the synthetic dataset size can increase significantly. However, this dataset may contain many low-quality candidates. In the next section, we will propose our method to filter out the low-quality candidates.

4.1.3 Synthetic data filtering

Our data filtering approach is based on sentence-level cosine similarity. This approach involves comparing the similarity between the original sentence and its corresponding back-translated sentence, enabling us to identify and eliminate sentence pairs that exhibit significant deviations from the original meaning. Our method distinguishes itself from Koehn's approach [36] by not requiring an effective acquisition of the linear mapping relationship between the embedding spaces of the source and target languages. Instead, we leverage a cosine similarity measure to assess the semantic similarity between sentences.

Data filtering based on cosine similarity An important aspect of this approach is sentence representation in different languages. Although multilingual LMs (e.g., XLM-RoBERTa) are possible to do that, the representations for out-of-the-box sentences are rather bad. Moreover, the vector spaces of different languages are not aligned, meaning that words or sentences with the same meaning in different languages are represented in different vectors. Reimers and Gurevych [39] proposed a straightforward technique to ensure consistent vector spaces across different languages. This method uses a PLM as a fixed Teacher model that produces good representation vectors of sentences. The Student model is designed to imitate the Teacher model. It means the same sentence should be represented as the same vector in the Teacher model and the Student one. To enable the Student model to work with additional languages, it is trained on parallel (translated) sentences. The translation of each sentence should also be mapped to the same vector as the original one.

In Figure 2, the Student model should map "Hello World" and the German translation "Hallo Welt" to the vector of Teacher("Hello World"). This is achieved by training the Student model using the mean squared error (MSE) loss.

Based on this approach, we first generate two bilingual datasets: Vietnamese-English and English-Khmer parallel sentence pairs from the original Km-Vi dataset, using the Google Translator API. This API is taken from the deep translator [4]. The Student model is then trained on both the Vietnamese-English dataset and the Khmer-English one to create semantic vectors for three languages: English, Vietnamese, and Khmer. The representation vector of a sentence is the average of the token embeddings based on the Student model. We calculate the sentence-level cosine similarity between each parallel in the synthetic parallel dataset and filter out pairs with low scores.

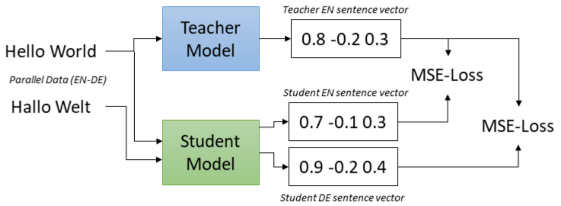


Figure 2: Given parallel data (e.g. English and German), train the student model such that the produced vectors for the English and German sentences are close to the teacher English sentence vector [39].

Data filtering using round-trip BLEU

The diagram of this method is represented in Figure 3. The process begins with the training of two NMT models: Km-Vi (source-to-target) and Vi-Km (target-to-source), using the given parallel sentence pairs. Next, we use the Vi-Km translation model to translate the monolingual sentences from the Vietnamese language to the Khmer one. We then take the translated sentences and back-translate them using the Km-Vi model. We evaluate the quality of sentence pairs based on sentence-level BLEU scores and discard sentence pairs with low scores.

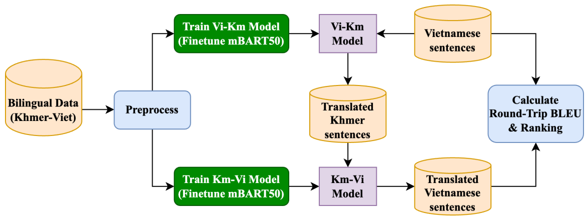

Bilingual Data (Khmer-Viet)
Preprocess
Train Vi-Km Model (Finetune mBART50)
Train Km-Vi Model (Finetune mBART50)
Vi-Km Model
Km-Vi Model
Vietnamese sentences
Translated Khmer sentences
Translated Vietnamese sentences
Calculate Round-Trip BLEU & Ranking

Figure 3: The diagram of the data filtering using round-trip BLEU.

4.2 Data augmentation method via english pivot language

A standard data augmentation method via English pivot language involves the translation of sentences in the target language from the original source-target parallel sentence pairs into English sentences. These English sentences are then translated into the source language to generate the source-target augmentation bilingual sentence pairs.

We propose an **"aligned"** version to improve the quality of the augmentation dataset. Given the original source-target sentence pair with a source sentence w_s and a target sentence w_t, we generate additional candidate sentences in the following way. The target sentence w_t is translated into

[4] https://github.com/nidhaloff/deep-translator

the source language using English pivot one. This step produces a candidate sentence in the source language w_{c1}. The target-to-source translation model described in Section 4.1 is used to generate another candidate sentence in the source language w_{c2}. The candidate pairs w_{c1} and w_{c2} are aggregated to get a temporary dataset. We carry out two filtering steps to remove low-quality parallel sentence pairs: (i) align parallel sentence pairs and (ii) data filtering. In the first step, the temporary dataset is aligned by three tools: Vecalign[5], Bleualign[6], and Hunalign [7]. Vecalign utilizes word embeddings to align sentences based on semantic similarity. Bleualign, on the other hand, uses the BLEU metric and n-gram overlap to align sentences in bilingual corpora. Hunalign is a heuristic-based tool that aligns parallel texts based on sentence length and lexical similarity. Sentence pairs that are aligned by two-third of the tools are selected to generate an aligned dataset. In the second step, the aligned dataset is filtered out based on the data filtering method in Section 4.1.3. As a result, we get an augmented dataset, which is combined with the synthetic parallel dataset from Section 4.1 and the original bilingual dataset to form the final training dataset.

5 Experiments

5.1 Experiment setup

We fine-tuned the mBART50 model on an RTX 3090 (24GB) GPU with different hyperparameters to choose the optimal parameter set for the model as follows: Adam optimization ($learning\_rate = 3\text{e}{-5}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1\text{e}{-8}$) with linear learning rate decay scheduling. The best set of hyperparameters is employed in all our experiments.

To evaluate the effectiveness of our experiments, we used the BLEU score [40] through sacreBLEU[8] - an implementation version to compute the BLEU score. A higher BLEU score indicates better translation quality.

5.2 Experimental scenarios

To evaluate the effectiveness of our proposed methods for low-resource NMT, we used the Km-Vi biligual dataset from Nguyen et al. [16]. This dataset consists of 142,000 parallel sentence pairs, which were divided into a training set of 140,000 sentence pairs and a test set of 2,000 ones. In order to prevent biased phenomena in experiments, Nguyen et al. [16] randomly selected 2,000 sentence pairs from the original bilingual dataset to form the test set, following the distribution ratio of domains and lengths.

Six scenario groups were carried out in our experiments.

Scenario group #1 - Baseline model: Fine-tune the mBART50 model on the original Km-Vi bilingual dataset (Scenario #1).

All scenario groups from #2 to #6 used additional bilingual datasets which were generated from the Vietnamese corpus or the Km-Vi original bilingual one. This dataset was combined with the original dataset to create a larger training corpus. The Vietnamese dataset were created by crawling from online news websites (i.e, vnexpress.net[9], dantri.com.vn[10]), then preprocess to remove noise and long sentences. The langdetect[11] library was used to filter out non-Vietnamese sentences.

Scenario group #2 (#2.1 to #2.6) - Combine Scenario #1 and Back-translation: To generate a synthetic parallel dataset, 120,000 sentences from the above mentioned Vietnamese dataset were selected using our data selection strategies. These sentences were then translated into the Khmer language by using our back-translation method. We implemented and compared four data selection methods and two decoding ones (i.e, sampling and beam search).

Scenario group #3 (#3.1 to #3.3) - Combine Scenario #2 and Data filtering: In this scenario, we compared two methods in the data filtering strategy: the Round-Trip BLEU [35] (#3.1) and our proposed sentence-level cosine similarity (#3.2) . We experimented with two types of data selection: TF-IDF (#3.1 and #3.2) and combination score(#3.3).

Scenario group #4 (#4.1 to #4.2) - Combine Scenario #1 and Data augmentation via English pivot language: We compared ”standard” and ”aligned” versions to generate an augmented dataset. The Google Translator API is used for the translation task.

Scenario group #5 (#5.1 to #5.2) - Combine Scenarios #3 and #4: We created a new training dataset through the best settings from **Scenarios #3** and **#4**.

Scenario group #6 (#6.1 to #6.2)- Combine Scenario #5 and Data Generation: In this experiment, at the back-translation step, each sentence from the Vietnamese dataset was translated into k corresponding Khmer candidate sentences. Then these sentences were filtered and combined with the original bilingual dataset to create a new training dataset.

6 Experimental results

This section presents a comprehensive evaluation of our system performance under various scenarios and compares the best results with other relevant research. The analysis of the augmented data’s quality is provided in Appendix 1.

6.1 Analysis our system performance using different scenarios

We evaluated our different scenarios on a test set with 2,000 parallel sentence pairs. The results of our scenarios are pre-

[5] https://github.com/thompsonb/vecalign

[6] https://github.com/rsennrich/Bleualign

[7] https://github.com/danielvarga/hunalign

[8] https://github.com/mjpost/sacrebleu

[9] https://vnexpress.net

[10] https://dantri.com.vn

[11] https://pypi.org/project/langdetect

Table 2: Experimental results

Scenario	Name	Data Augmentation Methods				BLEU (%)
		Back-translation			via English pivot language	
		Data Selection	Decoding Strategy	Data Filtering		
#1	Baseline model	-	-	-	-	52.32
#2.1	#1 + Back-translation	Randomness	Beam search	-	-	53.16
#2.2	#1 + Back-translation	Randomness	Sampling	-	-	53.49
#2.3	#1 + Back-translation	TF-IDF	Beam search	-	-	53.83
#2.4	#1 + Back-translation	TF-IDF	Sampling	-	-	53.96
#2.5	#1 + Back-translation	Cosine similarity	Sampling	-	-	53.98
#2.6	**#1 + Back-translation**	**Combination Score**	**Sampling**	-	-	**54.08**
#3.1	#2 + Data Filtering	TF-IDF	Sampling	Round-Trip BLEU	-	54.27
#3.2	#2 + Data Filtering	TF-IDF	Sampling	Cosine similarity	-	54.38
#3.3	**#2 + Data Filtering**	**Combination Score**	**Sampling**	**Cosine similarity**	-	**54.48**
#4.1	#1 + Data Augmentation	-	-	-	Standard	52.98
#4.2	**#1 + Data Augmentation**	-	-	-	**Aligned**	**53.29**
#5.1	#3 + #4	TF-IDF	Sampling	Cosine similarity	Standard	54.51
#5.2	**#3 + #4**	**Combination Score**	**Sampling**	**Cosine similarity**	**Aligned**	**54.93**
#6.1	#5 + Data Generation	Combination Score	Sampling	Cosine similarity	Standard	55.13
#6.2	**#5 + Data Generation**	**Combination Score**	**Sampling**	**Cosine similarity**	**Aligned**	**55.37**

sented in Table 2. The baseline **Scenario #1** achieved a 52.32% BLEU score.

Scenario group #2 shows that the combination score gave the best results and the sampling decoding method is better than the beam search method.

Table 3: Effect of BLEU filtering threshold in the data filtering using round-trip BLEU in the **scenario #3.**

Scenario	Threshold	BLEU (%)
#3	10	54.02
#3	**15**	**54.27**
#3	20	54.16
#3	25	53.80

For **scenario groups #3**, first, we evaluated the effect of data filtering thresholds to the system's performance. Tables 3 and 4 show that the BLEU score increases when the filter threshold is increased, but up to a certain threshold, and then reduced. This means that as the filter thresholds increase, we can filter out more low-quality parallel sentence pairs in the synthetic bilingual dataset, but the size of this dataset decreases. The best thresholds were then applied for all scenarios in groups #3 in order to compare the system performance with other scenarios in Table 2.

Scenario #4 First, in the standard version, we evaluated the model's performance with different augmented sizes. The original bilingual dataset was combined with 30000, 50000, and 70000 augmented sentence pairs created by the data augmentation via the English pivot language to form three training datasets. The obtained BLEU scores gradually increased from 52.48%, 52.52%, to 52.98%, proportional to the enhanced data size. The best result using 70000 augmented sentence pairs was used to compare with other scenarios in Table 2 (Scenario #4.1). Scenario #4.2 also used 70000 augmented sentence pairs in the aligned version.

Table 4: Effect of the cosine filtering threshold in the data filtering using sentence-level cosine similarity in the **scenario #3**.

Scenario	Threshold	BLEU (%)
#3	0.5	54.02
#3	0.6	54.36
#3	**0.7**	**54.38**
#3	0.8	53.92

With a result of 54.93% BLEU score, **Scenario #5** shown the effectiveness when combined the best synthetic parallel datasets from Scenario #3 and 30,000 pair sentences augmented in Scenario #4.

Finally, **Scenario #6**, we incorporated Scenario #5 with our generation strategy to get 55.37% BLEU points, which improved 3.05% BLEU scores compared to the baseline model. The results shown that the process of generating a synthetic dataset based on only one candidate with the highest probability was not enough. Taking k candidates and evaluating them helped us to retain more suitable candidates.

6.2 Comparison with other models

In addition to our scenario results above, we compared our best result with some models: Google Translator[12], pretrained multilingual seq2seq models, including mBART50 [37], m2m100-1.2B [41], and nllb-$*$ [42]-a multilingual translation model introduced by the Facebook AI[13] recently. The results shown in Table 5 indicated that our best model achieved best results for translating from the Khmer language to the Vietnamese one. In addition, our current

[12] https://github.com/nidhaloff/deep-translator

[13] https://ai.facebook.com/

approach had a better performance than our previous model [18] with 0.86% BLEU score higher.

Table 5: Comparison our system results to other models

Models	BLEU (%)
facebook/mbart50	12.74
facebook/m2m100-1.2B	22.44
facebook/nllb-200-distilled-600M	32.48
facebook/nllb-200-distilled-1.3B	36.51
facebook/nllb-200-3.3B	37.81
Google Translator	50.07
Our previous work [18]	54.51
Our best model	**55.37**

7 Conclusions

This research presents an approach to address the low-resource challenge in Khmer-Vietnamese NMT. The proposed method utilizes the pretrained multilingual model mBART as the foundation for the MT system, complemented by various data augmentation strategies to enhance system performance. These augmentation strategies encompass back-translation, data augmentation through an English pivot language, and synthetic data generation. The highest performance is achieved when combining the aforementioned augmentation methods with effective data selection and data filtering strategies, resulting in a significant 3.05% increase in BLUE score compared to the baseline model utilizing mBART with the original dataset. Our proposed approach outperforms the Google Translator model by 5.3% BLEU score on a test set of 2,000 Khmer-Vietnamese sentence pairs. Future work involves applying our proposed approach to other low-resource language pairs to demonstrate its generalizability.

References

[1] T. Khanna, J. N. Washington, and et al. Recent advances in apertium, a free/open-source rule-based machine translation platform for low-resource languages. *Machine Translation*, Dec 2021. https://doi.org/10.1007/s10590-021-09260-6.

[2] P. Koehn, F. J. Och, and et al. Statistical phrase-based translation. In *In Proceedings of NAACL*, page 48–54, 2003. https://doi.org/10.3115/1073445.1073462.

[3] P. Koehn, H. Hoang, and et al. Moses: Open source toolkit for statistical machine translation. pages 177–180. Association for Computational Linguistics, 2007. https://doi.org/10.3115/1557769.1557821.

[4] K. Cho, B. Merriënboer, and et al. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of EMNLP*, pages 103–111, 2014. https://doi.org/10.3115/v1/w14-4012.

[5] D. Suleiman, W. Etaiwi, and A. Awajan. Recurrent neural network techniques: Emphasis on use in neural machine translation. In *Informatica*, 2021. https://doi.org/10.31449/inf.v45i7.3743.

[6] Y. Tian, S. Khanna, and A. Pljonkin. Research on machine translation of deep neural network learning model based on ontology. In *Informatica*, 2021. https://doi.org/10.31449/inf.v45i5.3559.

[7] S. Edunov, M. Ott, and et al. Understanding back-translation at scale. In *Proceedings of EMNLP*, pages 489–500, 2018. https://doi.org/10.18653/v1/d18-1045.

[8] A. Vaswani, N. Shazeer, and et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. https://doi.org/10.48550/arXiv.1706.03762.

[9] Y. Chen, Y. Liu, and et al. A teacher-student framework for zero-resource neural machine translation. In *Proceedings of ACL (Volume 1: Long Papers)*, pages 1925–1935, 2017. https://doi.org/10.18653/v1/p17-1176.

[10] Y. Kim, P. Petrov, and et al. Pivot-based transfer learning for neural machine translation between non-English languages. In *Proceedings of EMNLP-IJCNLP*, pages 866–876, 2019. https://doi.org/10.18653/v1/d19-1080.

[11] R. Sennrich, B. Haddow, and et al. Improving neural machine translation models with monolingual data. In *Proceedings of ACL (Volume 1: Long Papers)*, pages 86–96, 2016. https://doi.org/10.18653/v1/p16-1009.

[12] J. Zhang and C. Zong. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of EMNLP*, pages 1535–1545, 2016. https://doi.org/10.18653/v1/d16-1160.

[13] L. Mike, L. Yinhan, and et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, 2020. https://doi.org/10.18653/v1/2020.acl-main.703.

[14] C. Raffel, N Shazeer, A. Roberts, and et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. https://doi.org/10.48550/arXiv.1910.10683.

[15] Y. Liu, J. Gu, N. Goyal, and et al. Multilingual denoising pre-training for neural machine translation. *Transactions of ACL*, 8:726–742, 2020. https://doi.org/10.1162/tacl_a_00343.

[16] Van-Vinh Nguyen, , Huong Le-Thanh, and et al. KC4MT: A high-quality corpus for multilingual machine translation. In *Proceedings of LREC*, page 5494–5502, 2022.

[17] N. H. Quan, N. T. Dat, N. H. M. Cong, and et al. ViNMT: Neural machine translation toolkit, 2021. https://doi.org/10.48550/arXiv.2112.15272.

[18] V.H Pham and Le T.H. Improving khmer-vietnamese machine translation with data augmentation methods. In *Proceedings of SoICT '22*, pages 276–282, 2022. https://doi.org/10.1145/3568562.3568646.

[19] J. Devlin, M. Chang, and et al. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL: Human Language Technologies*, pages 4171–4186, 2019. http://doi.org/10.18653/v1/n19-1423.

[20] J. Zhu, Y. Xia, L Wu, and et al. Incorporating bert into neural machine translation, 2020. https://openreview.net/forum?id=Hyl7ygStwB.

[21] S. Rothe, S. Narayan, and et al. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of ACL*, 8:264–280, 2020. https://doi.org/10.1162/tacl_a_00313.

[22] B. Zoph, D. Yuret, and et al. Transfer learning for low-resource neural machine translation. In *Proceedings of EMNLP*, pages 1568–1575, 2016. https://doi.org/10.18653/v1/d16-1163.

[23] J. Hu, L. Zhang, and D. Yu. Improved neural machine translation with paraphrase-based synthetic data. In *Proceedings of NAACL*, 2019.

[24] X. Niu and et al. Subword-level word-interleaving data augmentation for neural machine translation. In *Proceedings of EMNLP*, 2018.

[25] Z. Liu and et al. Word deletion data augmentation for low-resource neural machine translation. In *Proceedings of ACL*, 2021.

[26] H. Wang and et al. Multi-objective data augmentation for low-resource neural machine translation. In *Proceedings of IJCAI*, 2019.

[27] C. Chu and et al. Domain adaptation for neural machine translation with limited resources. In *Proceedings of EMNLP*, 2020.

[28] M. Johnson, M. Schuster, and et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of ACL*, 5:339–351, 2017. https://doi.org/10.1162/tacl_a_00065.

[29] R. C. Moore and W. Lewis. Intelligent selection of language model training data. pages 220–224. Proceedings of ACL, 2010. https://aclanthology.org/P10-2041.

[30] M. Wees, A. Bisazza, and et al. Dynamic data selection for neural machine translation. In *Proceedings of EMNLP*, pages 1400–1410, 2017. https://doi.org/10.48550/arXiv.1708.00712.

[31] R. Wang, A. Finch, and et al. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of ACL*, pages 560–566, 2017. https://doi.org/10.18653/v1/p17-2089.

[32] S. Zhang and D. Xiong. Sentence weighting for neural machine translation domain adaptation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3181–3190, August 2018. https://aclanthology.org/C18-1269.

[33] C. C. Silva, C. Liu, and et al. Extracting in-domain training corpora for neural machine translation using data selection methods. In *Proceedings of the Third Conference on Machine Translation*, pages 224–231, 2018. https://doi.org/10.18653/v1/w18-6323.

[34] A. Poncelas and et al. Data selection with feature decay algorithms using an approximated target side. 2018. https://doi.org/10.48550/arXiv.1811.03039.

[35] A. Imankulova, T. Sato, and et al. Improving low-resource neural machine translation with filtered pseudo-parallel corpus. pages 70–78. Asian Federation of Natural Language Processing, 2017. https://aclanthology.org/W17-5704.

[36] P. Koehn, H. Khayrallah, and et al. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation*, pages 726–739, 2018. http://doi.org/10.18653/v1/w18-6453.

[37] Y. Tang, C. Tran, X. Li, and et al. Multilingual translation with extensible multilingual pretraining and finetuning, 2020. https://doi.org/10.48550/arXiv.2008.00401.

[38] J. Cho, E. Jung, and et al. Improving bi-encoder document ranking models with two rankers and multi-teacher distillation. In *Proceedings of SIGIR '21*, page 2192–2196, 2021. https://doi.org/10.1145/3404835.3463076.

[39] N. Reimers and I. Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation, 2020. https://doi.org/10.48550/arXiv.2004.09813.

[40] K. Papineni, S. Roukos, and et al. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318, 2002. http://doi.org/10.3115/1073083.1073135.

[41] A. Fan, S. Bhosale, H. Schwenk, and et al. Beyond english-centric multilingual machine translation, 2020. https://doi.org/10.48550/arXiv.2010.11125.

[42] NLLB Team. No language left behind: Scaling human-centered machine translation, 2022. https://doi.org/10.48550/arXiv.2207.04672.

A Appendix 1

To assess the quality of the augmented data, we present exemplary outputs from two methods described in Section 4 in Tables 6 and 7.

Table 6 exhibits examples generated by the back-translation method. Vi-Km sentence pairs in the first and second columns are added to the augmented training dataset if they pass the synthetic data filtering step. The table reveals that the NMT models employed in the back-translation process may still produce semantically incorrect sentences, particularly when translating proper names. Such sentences are subsequently filtered out during the data filtering process. Notably, no sentence pairs in the augmented dataset by this method exhibit poor quality.

In Table 7, we present examples of data augmentation via the English pivot language. Due to the relatively high quality of Google Translator, the augmented Vi-Km sentence pairs demonstrate a relatively high quality when the original Km-Vi sentence pair possesses good quality. However, discrepancies arise when the original Vietnamese-Khmer sentence pairs do not maintain complete semantic equivalence, leading to a similar outcome for the newly generated Khmer sentence and the original Vietnamese one. Consequently, in such instances, the data filtering step excludes the incorporation of the new sentence pair into the augmented dataset.

Table 6: Output examples of synthetic data generation process

Vi sentence (Vietnamese Monolingual Dataset)	**Km sentence generated by the Vi-Km model**	**Vi sentence generated by the Km-Vi model**	**Action**
Việt Nam đã thâm nhập và mở rộng thương mại tại thị trường này./*Vietnam has penetrated and expanded its trade in this market.*	វៀតណាម បាន ជ្រៀត ចូល និង ពង្រីក ពាណិជ្ជកម្ម ក្នុង ទីផ្សារ នេះ ។/*Vietnam has penetrated and expanded trade in this market.*	Việt Nam đã tham gia và mở rộng thương mại tại thị trường này./*Vietnam has joined and expanded trade in this market.*	Keep
Đoàn đại biểu kiều bào đã đến dâng hương ở tượng đài Vua Lê./*The overseas Vietnamese delegation came to offer incense at the statue of King Le.*	គណៈប្រតិភូ មកពី ប្រទេស វៀតណាម បាន មក បួងសួង នៅ រូបសំណាក ស្តេច Li Lei ។/*A delegation from Vietnam came to pray in the statue of King Li Lei.*	Một phái đoàn từ Việt Nam đã đến thăm các khu vực của Hoàng gia Li Lei./*A delegation from Vietnam visited the areas of Royal Li Lei.*	Filter out
Theo đó, các dụng cụ này dao động mức từ vài chục cho đến hàng chục triệu đồng./*Accordingly, these tools range from a few tens to tens of millions of dong.*	តាមរយៈ នេះ ឧបករណ៍ ទាំងនេះ មាន តម្លៃ ពី មួយ ដង ទៅ មួយ ដង ទៅ មួយ ដង ។/*Through this, these devices are priced from time to time.*	Bằng cách này, những thiết bị này có giá trị một lần, một lần, một lần./*This way, these devices are worth it once, once, once.*	Filter out

Table 7: Output examples of data augmentation process via english pivot language

Original Km sentence	**Original Vi sentence**	**Km augmented sentence**	**Action**
កញ្ចប់ ទិន្នន័យ ត្រូវបាន តម្រៀប តាម តំបន់ សម្រាប់ អ្នកទិញ ងាយស្រួល ជ្រើសរើស ។/*The data packets are sorted by area for the buyer to easily select.*	Các gói data được chia ra theo khu vực để người mua dễ dàng lựa chọn./*The data packages are divided by region for buyers to easily choose*	កញ្ចប់ ទិន្នន័យ ត្រូវបាន ចាត់ ថ្នាក់ តាម តំបន់ ដើម្បី ងាយស្រួល ជ្រើសរើស អ្នកទិញ ។/*Data packages are categorized by area for easy selection of buyers.*	Keep
មនុស្ស ប្រមាណ ៥០០ លាន នាក់ អាច ប្រឈម នឹង ភាពក្រីក្រ ដោយសារ វិបត្តិ សេដ្ឋកិច្ច ដ៏ ធ្ងន់ធ្ងរ បំផុត តាំងពី មុន មក ។/*An estimated 500 million people could face poverty due to the worst economic crisis ever.*	Thế giới đang đối mặt với cuộc suy thoái kinh tế sâu sắc nhất, được đánh giá là nghiêm trọng hơn các cuộc khủng hoảng trước đây./*The world is facing the deepest economic recession, which is considered to be more severe than previous crises.*	ពិភពលោក កំពុង ប្រឈមមុខ នឹង វិបត្តិ សេដ្ឋកិច្ច ដ៏ ជ្រៅ បំផុត ដែល ត្រូវបាន គេ ចាត់ទុក ថា ធ្ងន់ធ្ងរ ជាង វិបត្តិ មុន ៗ/*The world is facing the deepest economic recession, which is considered to be more severe than previous crises.*	Filter out

A Hybrid Deep Learning Approach to Keyword Spotting in Vietnamese Stele Images

Anna Scius-Bertrand[1], Marc Bui[2] and Andreas Fischer[1]
[1]University of Fribourg and HES-SO, Fribourg, Switzerland
[2]Ecole Pratique des Hautes Etudes, Paris, France
E-mail: anna.scius-bertrand@unifr.ch, marc.bui@ephe.psl.eu, andreas.fischer@unifr.ch

Keywords: hybrid deep learning, document image analysis, keyword spotting, annotation-free, Hausdorff edit distance, Vietnamese steles, Chu Nom

Received: April 5, 2023

In order to access the rich cultural heritage conveyed in Vietnamese steles, automatic reading of stone engravings would be a great support for historians, who are analyzing tens of thousands of stele images. Approaching the challenging problem with deep learning alone is difficult because the data-driven models require large representative datasets with expert human annotations, which are not available for the steles and costly to obtain. In this article, we present a hybrid approach to spot keywords in stele images that combines data-driven deep learning with knowledge-based structural modeling and matching of Chu Nom characters. The main advantage of the proposed method is that it is annotation-free, i.e. no human data annotation is required. In an experimental evaluation, we demonstrate that keywords can be successfully spotted with a mean average precision of more than 70% *when a single engraving style is considered.*

Povzetek: Predstavljen je hibridni pristop za iskanje ključnih besed v slikah z nagrobnikov, ki združuje globoko učenje in strukturno modeliranje Chu Nom znakov. Ključne besede so uspešno prepoznane s povprečno natančnostjo več kot 70%.

1 Introduction

Vietnamese steles are of great value for historians, as the stone engravings are a unique source of information to understand the social, economic, and belief structures in the villages. The Vietnamica[1] project, in particular, aims to investigate pious donations from ordinary people offered to local shrines. For this purpose, about 40,000 digital stele images are studied, which may contain hundreds of Chu Nom characters on a single image. To cope with this vast amount of characters, automatic image analysis methods are needed that are able to transcribe the contents of the steles into machine-readable form for searching and browsing. Some examples of stele images are shown in Figure 1, highlighting significant differences in column layout and image quality across different steles.

Although the state of the art for handwriting recognition for historical documents has made great progress in the past decades, it remains a difficult problem and an active field of research [1]. Keyword spotting [2] has been proposed early on as an alternative to automatic transcription for difficult cases, where a full transcription is not feasible with high accuracy. The goal is to identify specific search terms of interest, either by providing a template image of the keyword (*query-by-example*) or by providing a textual representation of the search term (*query-by-string*).

Similar to developments in other related fields, such as computer vision and natural language processing, the different methods for keyword spotting can be divided into three groups, namely *heuristic* methods, *machine learning based* methods, and *deep learning based* methods. They are roughly ordered by time, heuristic methods being the oldest, but they still coexist and new approaches are being developed for all three groups.

Heuristic methods incorporate domain knowledge about the handwriting and are able to match images directly, i.e. keyword images and images from the manuscript, in order to retrieve similar instances in a query-by-example scenario. Early examples include dynamic time warping methods based on contour features [3] and gradient features [4], as well as segmentation-free methods based on scale-invariant feature transform (SIFT) [5]. More recently, a graph-based approach has been proposed in [6], which relies on a structural representation of the handwriting and uses an approximate graph edit distance to compare handwriting graphs.

Machine learning methods pursue the paradigm of learning by example. They train keyword models with the help of learning samples, i.e. manually annotated handwriting images. In a first step, characteristic features are manually defined based on domain knowledge and in a second step, different machine learning methods are used to learn keyword models based on the features. Examples of this group of keyword spotting methods include methods based on hidden Markov models (HMM) with geometric

[1] https://vietnamica.hypotheses.org

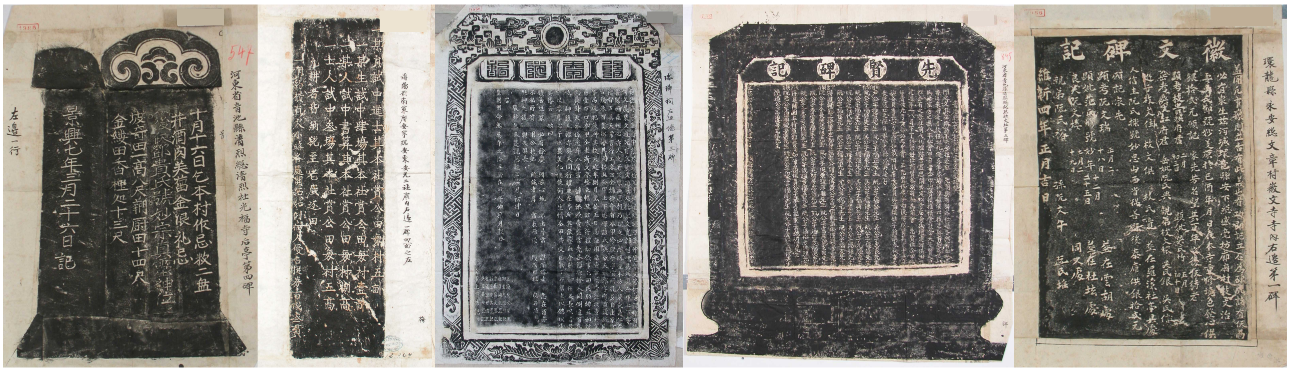


Figure 1: Example stele images.

features [7] and bag of local features [8], as well as bidirectional long short-term memory networks (BLSTM) with geometric features [9]. After training keyword models, the user can perform a query-by-string search, without the need of providing example images of the keyword.

Finally, deep learning methods are also based on the learning-by-example paradigm but they do not require manually defined features. Instead, they aim at learning characteristic representations, so-called embedding spaces, automatically from the data. Images as well as textual representations can be embedded in the same space, such that both query-by-example and query-by-string can be realized. A prominent example is the PHOCNet [10], which learns an embedding space based on pyramidal histogram of characters (PHOC) representations.

Today, the best keyword spotting performance is achieved by means of deep learning methods. However, they require a considerable amount of manually annotated training samples. In the case of historical Vietnamese steles, such learning samples can only be provided by experts, who are able to read the ancient Chu Nom script. It is thus time-consuming and costly to build a comprehensive training dataset, which is representative for the heterogeneous collection of stele images (see Figure 1). At the time of the writing, such a training set is not available for the 40,000 stele images.

In this article, we present a hybrid deep learning method for keyword spotting in historical Vietnamese stele images. It aims to combine deep learning with heuristic methods, such that the domain knowledge of the heuristic methods can compensate the lack of annotated learning samples. Indeed, it is an *annotation-free* method that does not require any human annotations at all.

The proposed method can be applied directly to the original stele images and consists of two processing steps. First, characters are detected using deep neural networks that are trained on synthetic images with printed Chu Nom characters and then auto-calibrated to real stele images. Secondly, the structure of the Chu Nom characters is modeled with a graph-based representation and matched with search terms using an approximate graph edit distance, in order to efficiently perform query-by-example keyword spotting. A comprehensive experimental evaluation is performed to measure the spotting performance.

The remainder of this article is structured as follows. Section 2 discusses related work on stele images, Section 3 provides more details about the content of the steles and the image acquisition, Section 4 presents the proposed keyword spotting method, and Section 5 details the experiments. Finally, conclusions are drawn in Section 6.

2 Related work

Initial work on the stele images has focused on the task of layout analysis with the aim to segment stele images into columns and characters. Such an initial segmentation is an important preprocessing step for character recognition. Notable work in this domain includes [11], where a heuristic method based on Voronoi diagrams is proposed to segment characters, and [12], where a deep learning approach based on semantic segmentation is pursued to detect columns with only a small number of human annotations. Recently, in [13] a deep learning method based on object detection networks has been introduced for character segmentation, which does not require human annotations and generalizes well to different layouts and engraving styles. In [14], a generative deep learning model has been suggested to create synthetic Chu Nom characters in different engraving styles. Recent work on Chu Nom also includes the U-Net based approach reported in [15], which was studied in the context of manuscripts.

Our method builds upon the character segmentation method of [13] and goes a step further to perform keyword spotting. For graph-based character modeling we rely on keypoint graphs, which have been studied for Latin scripts [6] and Chu Nom characters in manuscripts [16] before, but not for stone engravings. Important adaptations to the logographic writing system include the use of super-resolution to better model small strokes that distinguish similar Chu Nom characters. The graphs are efficiently matched using the Hausdorff edit distance [17], an approximation of graph edit distance that can be calculated in quadratic time with respect to the number of graph nodes.

Efficient graph matching is especially important in the context of super-resolution when Chu Nom graphs may contain over 100 nodes.

Preliminary results have already been published in a conference paper [18]. In this article, we provide a more detailed description of the hybrid deep learning method and significantly extend the experimental evaluation. Instead of considering only 8 steles, we conduct more comprehensive experiments on 20 stele images with manual ground truth. Furthermore, we study the important case of spotting keywords within the same style of engravings and compare it to a scenario with mixed styles. This study serves the purpose to better understand the possibilities and limitations of the proposed method.

3 Dataset

The 40,000 stele images represent about 25,000 steles, i.e. man-sized stones with engravings, which were erected in Vietnamese villages between the 16th and the 20th century. The majority of the steles represent donations made by villagers to the local shrines and are engraved in the ancient Vietnamese Chu Nom writing system [19]. However, they can also contain information about finances, constructions, and demarcations, thus informing about the social, economic, and religious life of the villages. The steles were erected for all to see and were able to withstand adverse weather conditions and armed conflicts. Nevertheless, they may contain degradations, fissures, and impacts, which may render parts of the steles illegible (see Figure 1).

The images of the steles were obtained by the French School of the Far East (EFEO) and the Institute of Han Nom Studies by means of stampings [19]. A sheet of paper is pressed on the stone and fixed with a binder, e.g. banana juice. Then, ink is applied with a roller on the paper over the entire surface of the stele, such that engravings appear in white and the stone background, as well as characters written in relief, appear dark in the color of the ink. Finally, the paper is photographed to obtain digital stele images. Such pictures of the stampings contain more character details and are easier to read when compared with pictures of the original steles.

In this article, we consider a research dataset of 20 stele images[2]. It encompasses all steles, for which we have obtained ground truth information so far at the level of individual characters, i.e. bounding boxes around the characters as well as their machine-readable Chu Nom transcription in unicode. Characters that are not readable are marked with a special symbol. In total, the dataset contains 5,138 characters, which corresponds to an average of about 257 characters per stele.

4 Hybrid deep learning

Figure 2 provides an overview of the proposed hybrid deep learning method for keyword spotting. At the core of the method is a deep learning model that is responsible to detect the location of main text characters on the stele images. It is trained on synthetic data and auto-calibrated to real data. Afterwards, a structural representation and comparison with keyword templates is performed for spotting.

The training data used for supervised learning of the deep learning model does not originate from human annotations. Instead, human knowledge is used to design the synthetic training data and to perform the auto-calibration. Also, human knowledge is used to model the characters with a graph-based representation and to perform the structural matching with the keyword templates based on heuristic methods.

In the following, the individual components are described in more detail.

4.1 Character detection

The deep learning model is an object detection network with a you look only once (YOLO) [20] architecture, which has originally been introduced for detecting objects in natural scenes, e.g. pedestrians in the context of autonomous driving. When applied to the problem of character detection on stele images, one of the main differences is that a large number of small objects need to be detected, rather than a small number of large objects. Therefore, it is important that the visual analysis is performed with a sufficiently high resolution, such that even small strokes of the logographic characters can be taken into account.

Two initial preprocessing steps are applied to the original images:

- Rescaling the stele images to a uniform height of 1024 pixels, while keeping the same aspect ratio.
- Inverting the colors, such that the engraved characters appear in dark color rather than white.

The height has been chosen to ensure that the stele images fit into the GPU memory. The resulting images are typically smaller than the originals. Inverting the colors has been chosen with respect to improved readability and more convenient generation of synthetic data.

The specific network architecture used for character detection is YOLOv5 [21], which analyzes the image at multiple sufficiently large scales to detect small characters on large steles. The *backbone* of YOLOv5 is a cross stage partial network (CSPNet) [22], which extracts convolutional feature maps. The *neck* is a path aggregation network (PANet) [23], which performs a combination of feature maps at different scales. Finally, a dual *head* is used to perform both classification and bounding box regression on the combined feature maps. Theoretically, the dual head would allow us not only to localize the characters but also

[2] The dataset is available at `https://github.com/asciusb/steles_kws_database`.

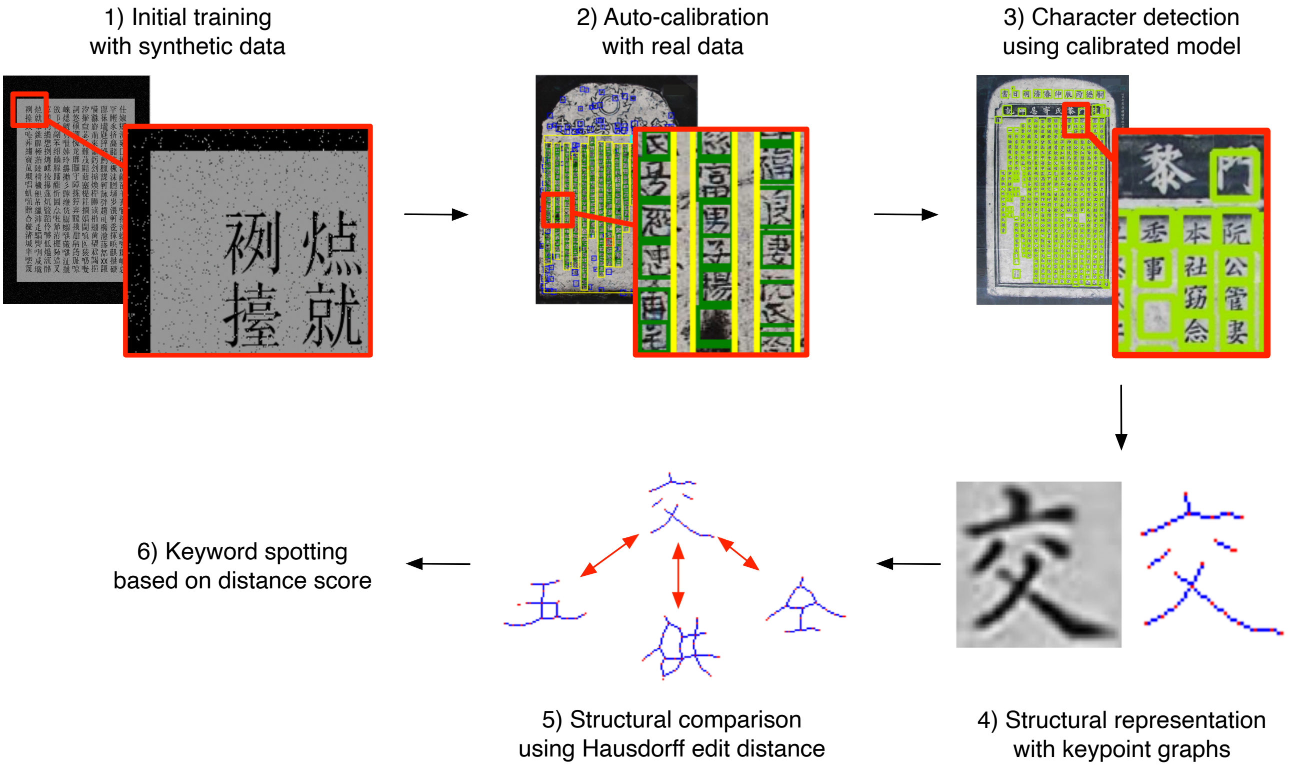


Figure 2: Overview of the hybrid deep learning method for keyword spotting. Green boxes represent detected characters and yellow boxes are detected columns.

to classify them. However, practical attempts to classification have failed when considering thousands of different Chu Nom character classes. Instead, the classification head only performs a binary classification, whether or not a character is present, and the regression head predicts the extent of the character bounding box.

4.2 Synthetic training

The initial training of the character detection network is based on thousands of fully synthetic training steles (Figure 2, step 1), for which the ground truth annotations, i.e. bounding boxes around the characters, is generated alongside with the synthetic images. The generation is guided by the following heuristic considerations:

- Color-inverted stele images contain dark character engravings on a gray stone background, surrounded by a black border.

- Characters are arranged in a column layout.

The data generation therefore proceeds as follows. First, a gray rectangle is drawn on a black background. Then, a Chu Nom font[3] is used to write random text on the gray rectangle in a random number of columns. Finally, random noise is added to the synthetic images by means of translation, blur, changes in brightness, as well as salt and pepper noise, in order to avoid overfitting of the character detection network.

4.3 Auto-calibration

After an initial training on synthetic data, the network is applied and adapted to real data by means of auto-calibration (Figure 2, step 2), following the procedure introduced in [13]. The aim of the auto-calibration is to replace the generic gray rectangle and black border of the synthetic training data with real stele backgrounds, such that the network can improve the separation between stele background and character foreground.

The following heuristic considerations are taken into account for detecting the main text area:

- Main text characters have approximately the same size and are organized in columns.

- The main text area is rectangular.

The auto-calibration is illustrated in more detail in Figure 3. After printing random Chu Nom text on simple backgrounds to create fully synthetic stele images, an initial training of the character detection network is performed. Afterwards, the network is applied to real stele images and layout analysis is used to recognize the main text area. Layout analysis consists of the following steps. The median

[3] The NomNaTongLight font available at http://www.nomfoundation.org.

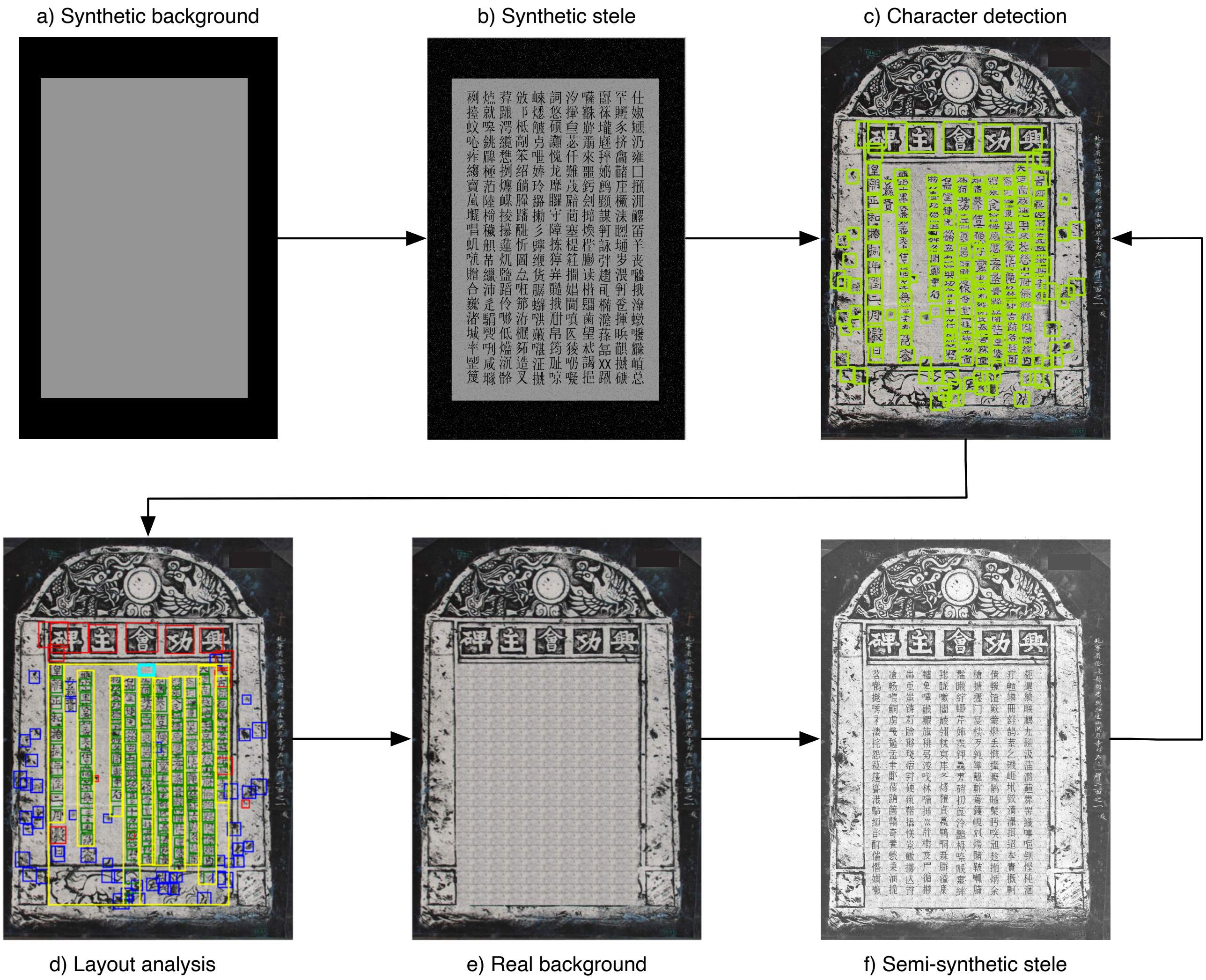


Figure 3: Auto-calibration of the deep learning model for character detection. Green boxes represent detected characters, red and blue boxes are characters discarded during layout analysis, yellow boxes are columns and the main text area, and the cyan box is a homogeneous background region.

box is calculated to estimate the size of the main text characters. Characters that are either too small (e.g. parts of ornaments or parts of the background), or too large (e.g. characters of the title above the main text) are discarded. Afterwards, unsupervised clustering using the DBSCAN [24] algorithm is performed to find the main text columns and thus the main text area around the columns (yellow rectangles in Figure 3). A homogeneous non-text region with low variance is determined as a pattern to fill the entire main text area (cyan rectangle in Figure 3). Finally, the Chu Nom font is used to write synthetic printed text on the main text area, similar to the generation of the initial training data, with the difference that a real stele background is present around the printed Chu Nom text.

The auto-calibration leads to new semi-synthetic training data, on which the initial network is further fine-tuned, thus adapting to real stele backgrounds and improving the detection accuracy. For further details, we refer to [13].

4.4 Structural representation

Once the characters have been detected by the calibrated network (Figure 2, step 3), the character images are modeled with a graph-based representation that captures their structure, i.e. the arrangement of individual strokes that constitute the character (Figure 2, step 4). We employ *keypoint graphs* [25], which have been used successfully for keyword spotting in the past for Latin manuscripts [6] as well as Chu Nom manuscripts [16] written with ink on parchment or paper.

The graph extraction is illustrated in Figure 4. First, a local text enhancement is applied by means of a difference of Gaussians (DoG) filter. Afterwards, the image is binarized with a global threshold and thinned to obtain strokes that have a width of one pixel. Endpoints and intersection points are added as nodes to the keypoint graph, labeled with their (x, y) coordinates. For circular structures, a random point is

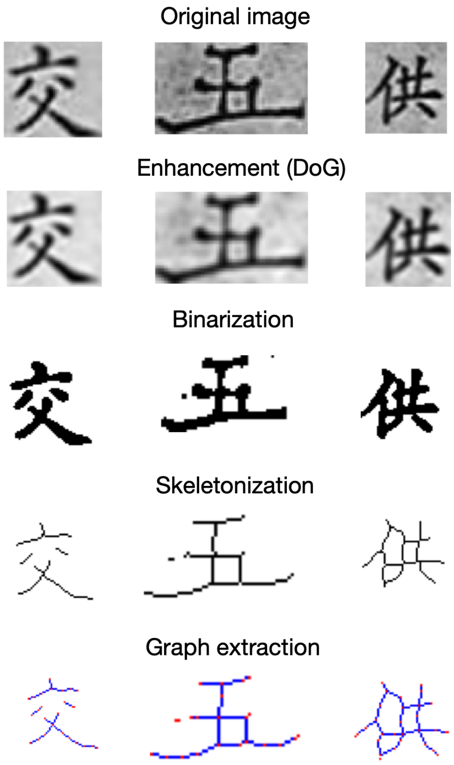


Figure 4: Keypoint graph extraction.

added as a node as well. To complete the initial set of nodes, additional points are added as nodes at regular intervals of D pixels on the skeleton image. Once all nodes have been added to the graph, their coordinate labels are normalized to zero mean and unit variance (z-score). Finally, neighboring nodes on the skeleton image are connected with unlabeled and undirected edges. For more details on the graph extraction, we refer to [26], whose implementation of keypoint graphs is used in the present article.

An important modification of the keypoint graph extraction, which led to successful spotting results on the stele images was to model the characters in super-resolution, in order to capture sufficient details about small strokes that may mark the difference between two similar Chu Nom engravings. To that end, the bounding box of the detection network is first translated back to the original image (inverting the downscaling to 1024 pixel height), then the character is cut out from the original image and upscaled to the same width S for all characters, while keeping the aspect ratio. When extracting graphs from a character image in super-resolution, i.e. when using values of S larger than the original width, it is possible that strokes in the keypoint graph contain more nodes than pixels in the original image. Hence, even very small strokes become more relevant in the graph-based representation.

4.5 Structural comparison

To compare the graphs of the character images with keyword graphs (Figure 2, step 5), we consider the graph edit distance [27, 28]. It is a general graph dissimilarity measure that is applicable to any kind of labeled graphs. With respect to a set of basic edit operations, typically insertion, deletion, and label substitution for nodes and edges, it calculates the minimum edit cost for transforming one graph into another. However, the exact graph edit distance is more of theoretical value than of practical relevance because it is an NP-complete problem, which can only be solved for small graphs with few nodes in reasonable time.

In order to cope with large character graphs, which may have over 100 nodes in super-resolution, we use the Hausdorff edit distance [17], an approximation of graph edit distance that calculates a lower bound in quadratic time. Derived from the Hausdorff distance between sets, it compares each node $u \in g_1$, plus its adjacent edges, of one graph g_1 with every node $v \in g_2$, plus its adjacent edges, of another graph g_2 and sums up the minimum edit cost $f(u, v)$ for matching the substructures. A special ϵ node is considered for insertions (ϵ, v) and deletions (u, ϵ). Formally,

$$\begin{aligned} HED(g_1, g_2) &= \sum_{u \in g_1} \min_{v \in g_2 \cup \{\epsilon\}} f(u, v) \\ &\quad + \sum_{v \in g_2} \min_{u \in g_1 \cup \{\epsilon\}} f(u, v) \end{aligned}$$

We use the Euclidean cost model for the structural comparison of keypoint graphs. It considers the Euclidean distance of the (x, y) labels for node label substitution, a constant cost c_n for node insertion and deletion, and a constant cost c_e for edge insertion and deletion.

4.6 Keyword spotting

To spot a Chu Nom character (Figure 2, step 6), n template images of the keyword are collected from real steles and keypoint graphs are extracted. Afterwards, the minimum HED score

$$score(g) = \min_{t \in T} HED(g, t)$$

is calculated for each character graph g of the stele images with respect to the template graphs $T = \{t_1, \ldots, t_n\}$. Finally, the character graphs are sorted according to the spotting score, such that the most similar character graphs appear in the top ranks.

For evaluating the spotting performance, precision (P) and recall (R) are calculated as

$$\begin{aligned} P &= \frac{TP}{TP + FP}, \\ R &= \frac{TP}{TP + FN}, \end{aligned}$$

with respect to the number of true positives (TP), false positives (FP), and false negatives (FN) for each possible score threshold. Then, the average precision (AP) is calculated for each keyword and the mean average precision (mAP)

$$mAP = \frac{1}{K}\sum_{i=1}^{K} AP_i$$

over all K keywords is used as the final performance measure for keyword spotting.

5 Experiments

5.1 Spotting scenarios

To evaluate the proposed hybrid deep learning method, the 5,138 ground truth characters of the 20 stele images (see Section 3) are randomly separated into three distinct sets for template selection (50%), validation (25%), and testing (25%), respectively.

The template selection set is used to select $n = 5$ templates per keyword, the validation set is used for optimizing hyper-parameters, and the test set is used for evaluating the final spotting performance. A total of $K = 128$ keywords are spotted, which appear at least 5 times in the template selection set, at least once in the validation set, and at least once in the test set.

We compare three spotting scenarios with respect to the use of human annotations, as listed in Table 1:

- The **fully annotated** scenario uses ground truth labels for parameter optimization as well as performance evaluation.
- The **font-validated** scenario does not require human annotations for parameter optimization. Instead, a synthetic font-based validation set is used (see below).
- The **annotation-free** scenario, which is the target scenario for the proposed method, does not require any human annotations. It evaluates the keyword spotting system with respect to automatically detected characters instead of ideal ground truth locations.

The synthetic font-based validation set is created as follows. 20 keywords are selected randomly and printed in 5 different Chu Nom fonts. 900 other characters are printed and added to the validation set, which is composed of 1,000 characters in total. Each of the keywords is then spotted on the validation set using a single template and the mAP results are used to compare and optimize different parameter settings.

Furthermore, we compare two spotting scenarios with respect to the engraving styles:

- The **same style** scenario spots keywords on each stele image separately, such that the engraving style of the keyword templates is the same as the style of the stele characters.
- The **mixed styles** scenario spots keywords across all 20 stele images, taking into account different engraving styles.

5.2 Parameter optimization

For the character detection network (see Sections 4.2 and 4.3), we consider only one set of hyper-parameters, i.e. the default parameters of the medium-sized YOLOv5m model[4]. The weights of the model are pretrained on the COCO [29] object detection dataset. The pretrained network is fine-tuned with 30,000 synthetic steles over 15 epochs until convergence. Afterwards, an additional fine-tuning epoch is used for auto-calibration with real stele backgrounds.

The parameters of the structural representation and the structural comparison (see Sections 4.4 and 4.5) are optimized in two steps. First, a default character width of $S = 150$ pixels and node distance of $D = 5$ pixels is fixed to evaluate a range of node and edge costs $c_n, c_e \in \{0.3, 0.6, 0.9, 1.2, 1.5, 1.8, 2.1\}$ on the validation set. Afterwards, the optimal node and edge costs are fixed and different character widths $S \in \{90, 120, 150, 180, 210\}$ and node distances $D \in \{3, 4, 5, 6, 7\}$ are evaluated on the validation set.

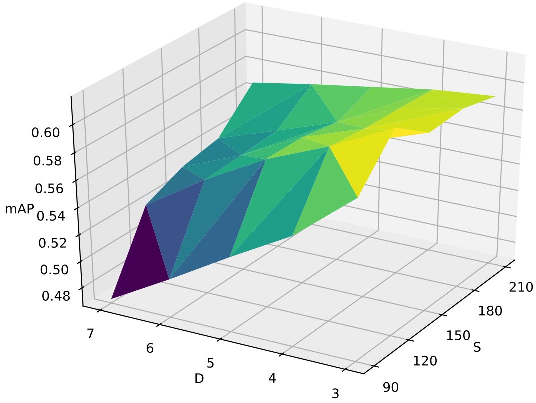


Figure 5: Optimization of structural representation for the fully annotated scenario.

The optimal parameters are listed in Table 2 for both the fully annotated and the font-validated scenario. When considering synthetic printed characters (font-validated), we observe that a larger node distance and a larger edge cost is preferred when compared with real stele characters (fully annotated). This may be due to the increased stability, i.e. less variability, of the character shapes and character background in the case of printed fonts.

A more detailed view on the optimization of S and D is provided in the three-dimensional visualizations in Figures 5 and 6. They show that changing the parameters lead to more significant differences in mAP for the real characters when compared with synthetic ones, indicating that the synthetic validation set may need to be improved to better represent the challenges encountered for real characters.

[4] github.com/ultralytics/yolov5, commit cc03c1d5727e178438e9f0ce0450fa6bdbbe1ea7

Table 1: Different keyword spotting scenarios with respect to human annotations.

	Parameter Optimization	**Performance Evaluation**
Fully annotated	Ground truth validation set	Ground truth test set
Font-validated	Synthetic font-based validation set	Ground truth test set
Annotation-free	Synthetic font-based validation set	Automatic detection test set

Table 2: Optimal parameters for structural representation and comparison.

Parameters	**Fully annotated**	**Font-validated**
Character width S	120	150
Node distance D	3	7
Node cost c_n	0.9	0.3
Edge cost c_e	0.9	2.1

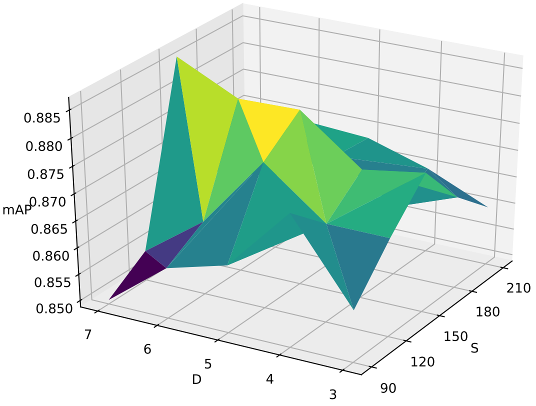


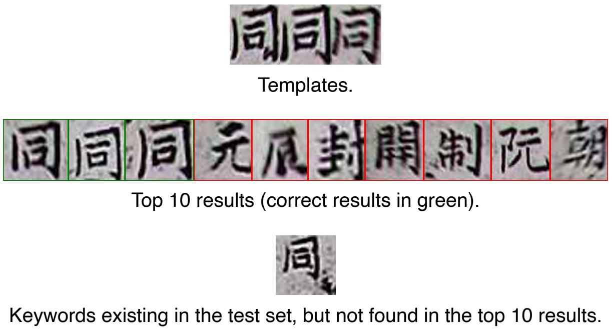
Figure 6: Optimization of structural representation for the font-validated scenario.

5.3 Runtime performance

For training the YOLO-based character detection model, we used 2 Nvidia Titan RTX GPUs. One training epoch on 30,000 stele images took 6.3 minutes on average and a total of 16 training epochs was sufficient for convergence. Detecting characters with the trained YOLO network took only a few milliseconds per stele.

For graph matching with the Hausdorff edit distance, we used computational nodes with 64 CPU cores (AMD EPYC, 2.25GHz). One graph comparison took 4.4 milliseconds on average, which allowed us to spot a single keyword template in about 1 second per stele.

Note that the graph comparisons need to be performed only once. Afterwards, the positions of all keywords in the collection of stele images can be indexed, such that historians can search and retrieve keywords quasi instantly based on the index.

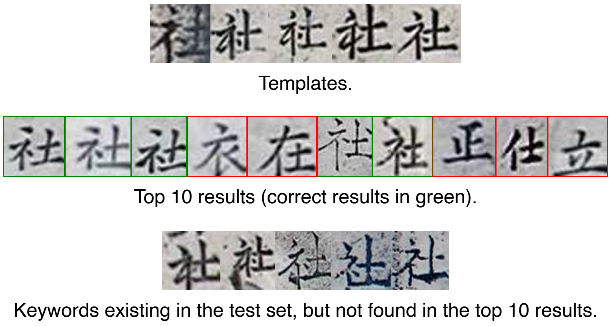


Templates.

Top 10 results (correct results in green).

Keywords existing in the test set, but not found in the top 10 results.

Figure 7: Qualitative spotting results for the same style spotting scenario.



Templates.

Top 10 results (correct results in green).

Keywords existing in the test set, but not found in the top 10 results.

Figure 8: Qualitative spotting results for the mixed styles spotting scenario.

5.4 Spotting performance

Table 3 shows the spotting performance on the test set when using optimized parameters. The results obtained for the steles are put into context with results obtained for Chu Nom manuscripts reported in [16]. These manuscripts are written with ink on parchment with a regular writing style, leading to better character detection quality and less noise in the character images when compared with the steles. The results are not directly comparable, because a different set of keywords was used, but they provide a point of reference for a less challenging keyword spotting task.

For the same style spotting scenario, the performance level is excellent with a mAP of 72% for the annotation-free scenario. The performance is similar to that of manuscripts, which typically are better readable and have less varia-

Table 3: Spotting performance in terms of mean average precision (mAP) on the test set.

	Fully annotated	**Font-validated**	**Annotation-Free**
Manuscripts (Kieu) [16]	0.76	0.78	0.77
Steles: Same style	0.85	0.81	0.72
Steles: Mixed styles	0.56	0.50	0.40

tions of the writing style when compared with stele images. However, the impact of optimizing the parameters on synthetic characters, rather than real ones, is stronger (mAP reduced from 85% to 81%) and the loss in mAP for automatic character detection is stronger as well (mAP reduced from 81% to 72%). These results show the increased difficulty of spotting Chu Nom characters in stele images and leave room for improvements regarding parameter optimization and character detection.

For the mixed styles spotting scenario, the performance drops significantly to a mAP of 40% for the annotation-free scenario. It indicates a limitation of the proposed hybrid deep learning method, which did not generalize well to engraving styles that are different from the keyword templates. An application to stele collections with similar engraving styles seems therefore more promising. Note, however, that we have only used 5,138 characters in our experiments. It is possible that the method will better generalize with a larger dataset.

Figures 7 and 8 provide qualitative spotting results for both scenarios. For same style spotting, 3 out of 4 characters are correctly spotted in the first 3 ranks. However, the fourth character is not part of the top 10 results, because of an error in automatic character detection, which has included some noise in the bottom right corner. For mixed style spotting, 5 out of 10 characters appear in the top 10 ranks, but not in the first 5 ranks. The remaining 5 characters are not part of the top 10 results, due to noise but also due to different engraving styles, which are not represented in the keyword template images.

6 Conclusions

The proposed hybrid deep learning approach to keyword spotting aims to combine the strengths of data-driven methods with knowledge-based modeling. In a first step, a deep convolutional neural network is trained on a large synthetic dataset to detect printed Chu Nom characters. By means of self-calibration, the network is then automatically adapted to the stele images. In a second step, the detected characters are modeled by means of keypoint graphs and the Hausdorff edit distance is used to efficiently perform a structural comparison for retrieving keywords.

Especially when the engraving style of the keyword is the same as the style of the stele characters, an excellent mean average performance of over 70% is achieved. In the case of mixed engraving styles, however, the spotting results drop to about 40% mean average precision. Although this performance level is still helpful for historians to browse large image collections of heterogeneous steles, there is clearly room for improvement.

There are several interesting lines of future research to further improve the results. Staying in the same style scenario, future work includes the investigation of style clustering, such that similar engraving styles can be identified across a large number of stele images. Noise removal methods may also be interesting to avoid spotting mistakes due to non-character artifacts.

With respect to the mixed style scenario, it may be necessary to perform some sort of data-driven learning to improve the spotting results, for example by means of geometric deep learning with graph neural networks [30]. In order to avoid the requirement of human annotations, it would be interesting to pursue a self-calibration strategy similar to the self-calibration of the character detection network.

Finally, a promising line of future research would be to generalize the proposed spotting method to other historical scripts and languages.

References

[1] Andreas Fischer, Marcus Liwicki, and Rolf Ingold, editors. *Handwritten Historical Document Analysis, Recognition, and Retrieval — State of the Art and Future Trends*. World Scientific, 2020.

[2] Raghavan Manmatha, C. Han, and E.M. Riseman. Word spotting: A new approach to indexing handwriting. In *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, pages 631—637, 1996.

[3] T. M. Rath and R. Manmatha. Word spotting for historical documents. *Int. Journal on Document Analysis and Recognition*, 9:139–152, 2007.

[4] K. Terasawa and Y. Tanaka. Slit style HOG features for document image word spotting. In *Proc. 10th Int. Conf. on Document Analysis and Recognition*, pages 116–120, 2009.

[5] M. Rusiñol, D. Aldavert, R. Toledo, and J. Lladós. Browsing heterogeneous document collections by a segmentation-free word spotting method. In *Proc. 11th Int. Conf. on Document Analysis and Recognition*, pages 63–67, 2011.

[6] Michael Stauffer, Andreas Fischer, and Kaspar Riesen. *Graph-based Keyword Spotting*. World Scientific, 2019.

[7] A. Fischer, A. Keller, V. Frinken, and H. Bunke. Lexicon-free handwritten word spotting using character HMMs. *Pattern Recognition Letters*, 33(7):934–942, 2012.

[8] L. Rothacker, M. Rusiñol, and G. A. Fink. Bag-of-features hmms for segmentation-free word spotting in handwritten documents. In *Proc. 12th Int. Conf. on Document Analysis and Recognition*, pages 1305–1309, 2013.

[9] V. Frinken, A. Fischer, R. Manmatha, and H. Bunke. A novel word spotting method based on recurrent neural networks. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(2):211–224, 2012.

[10] Sebastian Sudholt and Gernot A Fink. Phocnet: A deep convolutional neural network for word spotting in handwritten documents. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 277–282. IEEE, 2016.

[11] Thai V. Hoang, Salvatore Tabbone, and Ngoc-Yen Pham. Extraction of nom text regions from stele images using area voronoi diagram. In *10th International Conference on Document Analysis and Recognition*, pages 921–925, 2009.

[12] Anna Scius-Bertrand, Lars Voegtlin, Michele Alberti, Andreas Fischer, and Marc Bui. Layout analysis and text column segmentation for historical vietnamese steles. In *Proceedings of the 5th International Workshop on Historical Document Imaging and Processing*, pages 84–89, 2019.

[13] Anna Scius-Bertrand, Michael Jungo, Beat Wolf, Andreas Fischer, and Marc Bui. Annotation-free character detection in historical Vietnamese stele images. In *Proc. 16th Int. Conf. on Document Analysis and Recognition (ICDAR)*, pages 432–447, 2021.

[14] Jonas Diesbach, Andreas Fischer, Marc Bui, and Anna Scius-Bertrand. Generating synthetic styled chu nom characters. In *Proc. 18th Int. Conf on Frontiers in Handwriting Recognition (ICFHR)*, 2022.

[15] Kha Cong Nguyen, Cuong Tuan Nguyen, and Masaki Nakagawa. Nom document digitalization by deep convolution neural networks. *Pattern Recognition Letters*, 133:8–16, 2020.

[16] Anna Scius-Bertrand, Linda Studer, Andreas Fischer, and Marc Bui. Annotation-free keyword spotting in historical vietnamese manuscripts using graph matching. In *IAPR Joint International Workshops on Statistical Techniques in Pattern Recognition (SPR 2022) and Structural and Syntactic Pattern Recognition (SSPR 2022) : S+SSPR*, 2022.

[17] A. Fischer, C. Y. Suen, V. Frinken, K. Riesen, and H. Bunke. Approximation of graph edit distance based on Hausdorff matching. *Pat. Rec.*, 48(2):331–343, 2015.

[18] A. Scius-Bertrand, A. Fischer, and M. Bui. Retrieving keywords in historical vietnamese stele images without human annotations. In *Proc. 11th Int. Symposium on Information and Communication Technology (SoICT)*, 2022.

[19] Philippe Papin. Aperçu sur le programme "publication de l'inventaire et du corpus complet des inscriptions sur stèles du viêt-nam". *Bulletin de l'École française d'Extrême-Orient*, 90(1):465–472, 2003.

[20] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[21] Glenn et al. Jocher. ultralytics/yolov5: v4.0 - nn.silu() activations, weights & biases logging, pytorch hub integration. DOI: 10.5281/zenodo.4418161, 2021.

[22] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. Cspnet: A new backbone that can enhance learning capability of cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 390–391, 2020.

[23] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018.

[24] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining*, pages 226–231, 1996.

[25] A. Fischer, K. Riesen, and H. Bunke. Graph similarity features for HMM-based handwriting recognition in historical documents. In *Proc. Int. Conf. on Frontiers in Handwriting Recognition*, pages 253–258, 2010.

[26] Paul Maergner, Vinaychandran Pondenkandath, Michele Alberti, Marcus Liwicki, Kaspar Riesen, Rolf Ingold, and Andreas Fischer. Combining graph edit distance and triplet networks for offline signature verification. *Pattern Recognition Letters*, 125:527–533, 2019.

[27] H. Bunke and G. Allermann. Inexact graph matching for structural pattern recognition. *Pattern Recognition Letters*, 1(4):245–253, 1983.

[28] A. Sanfeliu and K. S. Fu. A distance measure between attributed relational graphs for pattern recognition. *IEEE Trans. on Systems, Man, and Cybernetics*, 13(3):353–363, 1983.

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[30] Pau Riba, Andreas Fischer, Josep Lladós, and Alicia Fornés. Learning graph edit distance by graph neural networks. *Pattern Recognition*, 120:108132, 2021.

Comparative Study of Missing Value Imputation Techniques on E-Commerce Product Ratings

Dimple Chehal, Parul Gupta, Payal Gulati, Tanisha Gupta
Department of Computer Engineering, J.C. Bose University of Science and Technology, YMCA, Faridabad, India
E-mail: dimplechehal@gmail.com, parulgupta_gem@yahoo.com, gulatipayal@yahoo.co.in,
tanishagupta067@gmail.com

Keywords: imputation techniques, missing value, recommender system, sparsity, user ratings

Received: May 5, 2022

Missing data is typical as it adds ambiguity to data interpretation, and missing values in a dataset represent loss of vital information. It is one of the most common data quality concerns, and missing values are typically expressed as NANs, blanks, or other placeholders. Missing values create imbalanced observations, biased estimates and sometimes lead to misleading results. As a result, to deliver an efficient and valid analysis, there arises a need to take the solutions into account appropriately. By filling in the missing values, a complete dataset can be created and the challenge of dealing with complex patterns of missingness can be avoided. In the present study, eight different imputation methods: SimpleImputer, KNN Imputation (KNN), Hot Deck, Linear Regression, MissForest, Random Forest Regression, DataWig, and Multivariate Imputation by Chained Equation (MICE) have been compared. The comparison has been performed on Amazon cell phone dataset based on three parameters: R- Squared Error (R^2), Mean Squared Error (MSE), and Mean Absolute Error (MAE). Based on the findings KNN had the best outcomes, while DataWig had the worst results for R- Squared error (R^2). In terms of Mean Squared Error (MSE) and Mean Absolute Error (MAE), the Hot Deck imputation approach fared best, whereas MissForest performed worst for Mean Absolute Error (MAE). The Hot Deck imputation approach seems to be of interest and should be investigated further in practice.

Povzetek: Primerjava tehnik imputiranja manjkajoče vrednosti pri ocenah izdelkov e-trgovine

1 Introduction

Missing data occurs frequently in research such as Clinical Trials, Climatology and Medicine as it adds a layer of ambiguity during data interpretation [9], [19], [1], [5]. Nowadays, most databases present a problem of incomplete data. Missing values in a dataset mean loss of important information. These are values that are not present in the data set and are written as NAN's, blanks, or any other placeholders. Missing value creates imbalanced observations, biased estimates and in some cases can direct to misleading results. There can be multiple reasons for the missing value in a dataset such as failure to capture data, incorrect measurements or defective equipment, data corruption, sample mishandling, a low signal-to-noise ratio, measurement inaccuracy, non-response, or a deleted anomalous result [15], [10]. Building a machine learning algorithm with a dataset containing missing values can have a major impact on machine learning models as well as on the outcomes. Missing values can be of both continuous and categorical types. To get more precise results, multiple techniques can be used to fill out missing values.

Many approaches for dealing with missing data have been presented in recent years, and they can be categorized as deletion and imputation. There are three common deletion approaches list wise deletion, pair-wise deletion, and feature deletion. The common approach in list wise or case elimination is to omit the cases with missing values and evaluate the remaining data. Pair-wise deletion, on the other hand, removes data only when the specific data points required to test a hypothesis are missing. The existing values are employed in statistical testing if there is missing data elsewhere in the data set. A pair-wise deletion maintains more information than a list wise deletion since it uses all information observed [11].

Imputation on the other hand is the process of identifying missing values and interchanging them with a substitute value is known as missing value imputation [13], [6]. The method of missing value imputation is depicted in Figure 1. The experiment begins with the selection of a dataset, which is then characterized as incomplete or complete based on the quantity of missing data in the dataset. When a dataset is classified as incomplete, it is split into two parts: complete data and missing data. Imputation methods employ the entire dataset to impute missing values in the dataset. After that, a complete dataset with no missing values is created. The performance of the imputation methods is computed when the whole dataset and experimental dataset are compared using performance measures.

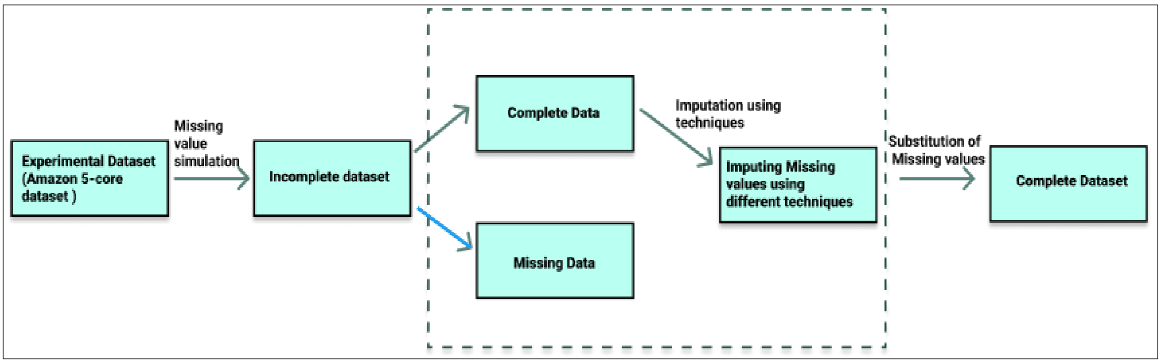

Experimental Dataset (Amazon 5-core dataset)
Missing value simulation
Incomplete dataset
Complete Data
Missing Data
Imputation using techniques
Imputing Missing values using different techniques
Substitution of Missing values
Complete Dataset

Figure 1: Missing value imputation process.

Single imputation and multiple imputations are two subgroups of the numerous imputation techniques. In single imputation, only one value is present for each missing cell and the value thus generated is used as the original value, although no imputation method can provide the exact value [18], [25]. The workflow for single imputation is depicted in Figure 2. First, the type of missing data is determined, and then single imputation is chosen from the two alternatives of single and multiple imputations, which is further separated into explicit and implicit modeling. The assumptions are explicit because the predicted distribution in explicit modeling is based on a formal statistical model, like multivariate normal. This process employs the mean imputation and regression imputation techniques. Hot Deck imputation, substitution, and cold deck imputation are all part of this procedure.

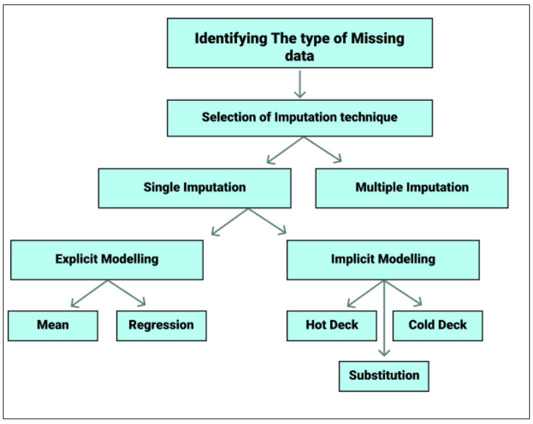


Figure 2: Work flow for single imputation.

In multiple imputations of a missing cell, multiple values are generated to impute the cell. Many complete data sets with various imputed values generate after which each data set is analyzed independently and the results computed. In contrast to single imputations, multiple imputations account for statistical uncertainty in the imputations [21], [7]. The workflow for multiple imputations is depicted in Figure 3. First, the type of missing data is determined, and then multiple imputations are chosen from the two alternatives of single and multiple imputations. Several imputations generates multiple values from separate imputed sets, which are then analyzed after calculating a single value for each missing value, and a single value is chosen from all the values to impute a missing value in the incomplete dataset. As a result, there are three separate phases to the multiple imputation technique:

a. M handles missing data, resulting in M complete data sets.
b. After that, the M full data sets are analyzed.
c. For the final imputation result, the outcomes of all M imputed data sets are pooled.

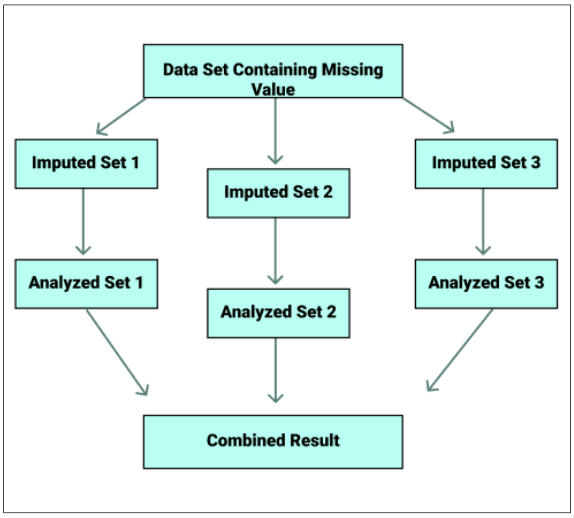


Figure 3: Work flow for multiple imputation.

Existing imputation techniques have been compared using R Squared (R^2), Mean Absolute Error (MAE), and Mean Squared Error (MSE) Metrics.

There are three main types of missing values:

a. Missing completely at random (MCAR)
b. Missing at random (MAR)
c. Not missing at random (NMAR)

The relationship between missingness and the values of the variables in the dataset is stated by the missing data mechanism. A dataset Y is stated to be a combination of a variable that is observed and a variable that is missing (Y_{obs} and Y_{mis}, respectively). The first type is known as missing completely at random (MCAR), in which the value itself or any known value is not a determinant of the missing values. Thus, Y_{obs} and Y_{mis} have no effect on the likelihood of a missing value [11], [8], [14]. The second type is Missing at random (MAR) is the polar opposite of MCAR, in which missing values are dependent on known values or on the value itself. Thus, the probability of a missing value is independent of Y_{mis} or Y_{obs}. MAR and MCAR can be ignored because it is impossible to adjust for the missingness. The final form Not missing at random (NMAR), where the probability of a missing occurrence varies [14].

This study is divided into seven sections. A brief past related work is provided in Section 2. Different missing value patterns are explained in section 3. In section 4, a description of the dataset as well as data analysis is given. The paper's results, as well as the evaluation criteria used, are explained in section 5, and the study's conclusion is shown in section 6.

2 Related work

There are multiple techniques to impute missing value's the first and the oldest one is SimpleImputer in which mean of a single column is computed to fill the missing value or cell with the mean computed of the rest of the cells of that column. SimpleImputer leads to poor imputation because it ignores correlation between different features [14].

Whenever the variables have a non-linear connection, linear regression-based imputation may underperform. The conditional model for imputation is Classification and Regression Trees (CART) [3]. Random forest extensions also have yielded encouraging results [22]. The decision tree-based imputation techniques are non-parametric algorithms that do not forecast the distribution of the data.

K-Nearest Neighbors (K-NN) based imputation is one of the most often used non-parametric techniques. This technique replaces the observed values in dimension d for each missing element with the mean of the K-nearest neighbors' d^{th} dimension [24]. Sequential K-NN is a K-NN extension that begins by imputing missing values from observations with the fewest missing dimensions and then moves on to the next unknown entries while reusing the previously imputed values [12]. Iterative K-NN uses an iterative procedure to re-estimate the estimates and select the closest neighbors based on the previous iteration's estimations.

Single imputation approaches produce a single set of finished data that may be utilized for statistical analysis. Whereas, multiple imputations, impute numerous times (each set may be different), then conduct statistical tests on all sets and combine the results. This strategy can capture the variability in missing data and, as a result, produce potentially more accurate estimates for the wider statistical problem. Multiple imputation approaches, on the other hand, are slower and necessitate pooling of results, which may not be appropriate for some applications.

The process for generating several estimates of missing data varies within the multiple imputation frameworks. A common multiple imputation method, multivariate imputation by chained equations (MICE), generates estimates using predictive mean matching, Bayesian linear regression, logistic regression, and other techniques [4]. Missing data imputation is still a hot topic in research because of its importance. Despite the fact that there are several approaches, many of them have serious shortcomings and their own pros.

In the event of missing values, information management is critical. Planning, organizing, structuring, processing, regulating, assessing, and reporting information operations are all part of the information management cycle. The major goal of information management is to produce and manage data in order to gain better insights; hence in missing value imputation, missing data is discovered using various strategies both of single imputation and multiple imputations in order to gain a better understanding of datasets and compute important and numerically significant conclusions. When managed information is fed into any algorithm, the algorithm's performance improves, ultimately assisting in the resolution of recent technological issues.

3 Missing data patterns and imputation approaches

Missing data patterns explain which values in the dataset are missing and which values should be observed. Univariate, monotone, and non- monotone missing data patterns are the three types of missing data patterns.

a. ***Univariate:*** When only one variable has missing data, the data is classified as univariate missing data pattern. To be classified in Univariate, the missing values should be in one column [17].

b. ***Monotone:*** When data is ordered and the pattern is frequently connected with longitudinal studies where participants drop and never return, it is called Monotone data. This is easier to detect because they are more visible and distinguishable [2].

c. ***Non- Monotone:*** Data is non-monotone when missing values in one variable or column have no effect on the values of other columns or the missing values of other columns [20].

Missing value imputation is the most important part of data analysis since it ensures that the dataset is complete and the results are computed correctly. There are mainly two types of imputation techniques single imputation and multiple imputations. In this experiment, techniques like SimpleImputer, KNN Imputation (KNN), Hot Deck, Linear Regression, MissForest, Random Forest Regression, DataWig, and Multivariate Imputation by Chained Equation (MICE) will be compared and evaluated. Advantages and disadvantages of these techniques have been shown in Table 1.

a. ***Imputation using SimpleImputer:*** SimpleImputer is a scikit-learn class that aids with missing data imputation in datasets used for predictive modeling [16], [23]. It substitutes a placeholder for the NaN values. SimpleImputer employs a variety of strategies to impute values, one of which is the use of mean/median to replace missing values. In this technique, the mean or median of the non-missing values is computed, and the missing values in the column are imputed using the computed mean or median value. This technique is best applied to numerical values rather than categorical ones. Mean imputation is quick and simple to implement, it preserves the mean of the observed data. This implies if data is Missing completely at random (MCAR), the estimate of mean remains unbiased. However, mean imputation is less accurate than other impute techniques.

b. ***Imputation using KNNImputer***: KNNImputer is a *scikit-learn* python machine learning library that aids in nearest neighbor imputation [16]. In KNN imputation, the distance between data points is measured and the number of contributing neighbors is chosen for each prediction. The number of nearest neighbors used to predict a missing value is usually controlled by the value of K, which has a direct impact on the KNN algorithm's performance. A high K value reduces the impact of random error on variance, but it also increases the risk of missing important small-scale patterns. When selecting an appropriate value of K, it is critical to strike a balance between over fitting and under fitting.

c. ***Hot Deck imputation***: In a sample set with similar values on all other variables, Hot Deck imputation selects one value at random from each individual set of values. This means that all records in the dataset with similar values in other variables are searched, and any one record is selected and utilized to impute the missing values [17].The benefit is that no outliers are created in the dataset as a result of this method.

d. ***Imputation using Linear Regression***: Regression is a two-step procedure in which a regression model is first constructed utilizing all of the available and complete data points. The created model is then used to impute missing data. In linear regression a regression equation is formed in which the best predictors are classed as independent variables, whereas variables with missing data are labeled as dependent variables. The missing values are predicted using a regression equation using independent and dependent variables. Values for the missing variable are inserted in an iterative procedure, and then all cases are utilized to forecast the dependent variable. These steps are repeated until the projected values are almost identical from one step to the next, at which point they converge.

e. ***Imputation using MissForest***: MissForest is a machine learning data imputation method that is based on the random forest algorithm [22]. Firstly the missing data are imputed using median/mode imputation. Then the non-missing values are marked as training rows and missing values are marked as predicted, the training rows are fed into a random forest model used to predict the missing values. The training rows are then fed into a random forest model that predicts missing values. The projected values are then imputed to replace the existing values, resulting in a dataset that is full and free of missing values. To enhance imputation in each iteration, the entire procedure is done numerous times. MissForest is capable of handling numerical, categorical, and mixed data types. MissForest is created with the *missingpy* library.

f. ***Imputation using Random Forest Regression***: The Random Forest is a Meta estimator technique that employs averaging to increase predicted accuracy and control over-fitting by fitting several classification decision trees on various sub-samples of the dataset. Random forest regression is a supervised learning approach for regression that uses the ensemble learning method. The ensemble learning method combines predictions from several machine learning algorithms to get a more accurate forecast than a single model. For regression problems, the mean or average forecast of the individual trees is computed known as aggregation. Instead of depending on individual decision trees, the main idea is to aggregate numerous decision trees to determine the final outcome. As a fundamental learning model, Random Forest uses several decision trees. Row and feature sampling are done at random from the dataset, resulting in sample datasets for each model this process is known as bootstrap.

g. ***Imputation using Deep Learning (DataWig)***:DataWig is a machine learning package that employs Deep Neural Networks to impute missing values in a dataset [2]. DataWig combines deep learning feature extraction with automatic hyper parameter tuning. This approach applies to both categorical and non-numerical data. DataWig first determines the type of each column. The column is then translated to a numerical representation. DataWig can be used to train on both the CPU and the GPU. DataWig typically works on a single column at a time, with the target column holding information about the imputing column supplied ahead of time.

h. ***Imputation using Multivariate Imputation by Chained Equation (MICE)***: In multiple imputations, many imputations are created for each missing value. It means filling the missing values multiple times and creating multiple complete datasets. One well-known algorithm for multiple imputations is Multiple Imputation by Chained Equation (MICE). MICE works under the assumption that missing data is Missing at random (MAR) or Missing completely at random (MCAR). Implementing MICE when data is not MAR could result in biased estimates. MICE is very flexible technique and can handle multiple variables and complexities of varying types at a time. It employs a divide-and-conquer strategy to impute missing values in dataset variables, focusing on one variable at a time. Once the emphasis is placed on that variable, it uses all of the other variables in the data set to forecast missingness in that variable. A regression model, the form of which is dictated by the nature of the focal variable, is used to make the prediction.

Table 1: Advantages and disadvantages of imputation techniques

S. No	Method	Advantages	Disadvantages
1.	SimpleImputer	1. It's a simple and quick procedure. 2. It's suitable for small numerical datasets.	1. Correlation between features is not taken into account. 2. Not extremely precise.
2.	KNNImputer	1. Better than SimpleImputer in terms of accuracy	1. KNN operates by memorizing the entire training dataset 2. Sensitive to outliers
3.	Hot Deck imputation	1. Because of residuals, the imputed data will have the same distribution shape as the actual data. 2. It's good for categorical data.	1. It's not good for small sample sizes.
4.	Linear Regression	1. For numeric data, this strategy is more effective.	1. If the prediction power is poor, this approach will perform poorly.
5.	MissForest	1. The looping over missing data point's process is repeated numerous times, with each iteration improving on improved data. 2. It can be used with both numerical and category data. 3. There is no need for preprocessing.	1. Time consuming because the number of iterations is dependent on the size of the dataset. 2. Expensive to operate MissForest
6.	Random Forest Regression	1. Outlier resistant. 2. Does a good job with non-linear data. 3. Less chance of over fitting. 4. Performs well on a huge dataset.	1. Slow and steady training. 2. Linear approaches with a lot of sparse features aren't recommended.
7.	Deep Learning (DataWig)	1. It works with categorical data. 2. Supports both CPUs and GPUs	1. Slow when dealing with large datasets 2. Imputation of a single column.
8.	Multivariate Imputation by Chained Equation (MICE):	1. Unbiased estimates, which are more reliable than ad hoc responses to missing data	1. MICE works under the assumption that missing data is Missing at random (MAR) or Missing completely at random (MCAR)

4 Experiments on rating predictions

This section details the dataset used and its corresponding analysis.

4.1 Dataset description

In this study, the publicly accessible dataset from Amazon of cell phone and accessories has been used. In the 5-core dataset, all users and items have at least five reviews. It consists of 1048570 rows and 12 columns. The 12 columns are overall (rating of product), Verified (for verified product by Amazon), ReviewTime (time of review submission), ReviewerID (ReviewerID of each reviewer), Asin (product ID), Style (sparse value pertaining to product's color), ReviewerName (name of the reviewer), ReviewText (review text), Summary (review summary), UnixReviewTime (review time (UNIX time)), Vote (total number of votes earned by a product), Image (product image link).

The primary columns to pay attention are verified, vote and rating. Then the dataset is preprocessed to ensure that every product has a vote value because the data is massive and sparse in the vote column. The dataset was reduced to 90714 rows and 12 columns after preprocessing.

4.2 Data analysis

The principle of analysis is depicted in Figure 4. Initially, there were no missing values in the dataset. As a result, missing values of about 4\% were created in the original dataset (Amazon 5-core) based on the MCAR model in the overall column, and imputation was performed using several strategies. These missing values were simulated and imputed using the eight techniques and three evaluation criteria (R-squared error, MAE, and MSE). R-squared, a statistical measure represents the degree of goodness of fit of a regression model. The best r-square value is 1.

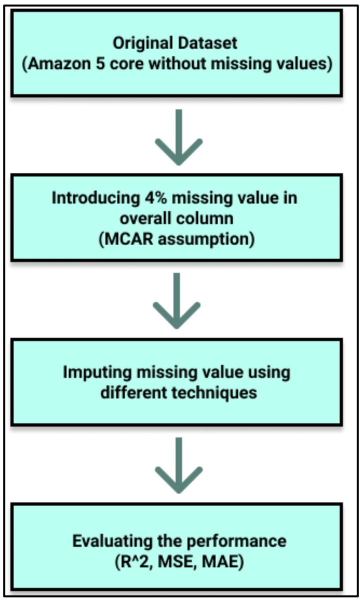

Original Dataset
(Amazon 5 core without missing values)
Introducing 4% missing value in overall column
(MCAR assumption)
Imputing missing value using different techniques
Evaluating the performance
(R^2, MSE, MAE)

Figure 4: Principal of analysis

5 Results and discussion

Missing values were imputed using eight distinct imputation approaches. With the use of the vote and verified columns, all of the strategies effectively imputed the missing values that were present in the Overall column. The three-assessment metrics were used to measure the performance of the techniques R^2, MSE and MAE, Table 2 compares all the eight approaches based on these assessment metrics.

a. ***R-squared:*** The closer the r-squared value is to 1, the better the model fits. When the fitted models are worse than the average fitted model, the R-Squared value can be negative. The R-squared is determined by dividing the sum of squares of residuals from the regression model (SS$_{\mathrm{RES}}$) by the total sum of squares of errors from the average model (SS$_{\mathrm{TOT}}$), then subtracting 1. The R-squared is mathematically defined by the equation 1:

$$R^2 = 1\text{-}\ \frac{SS_{RES}}{SS_{TOT}} \quad = 1\text{-}\ \frac{\sum_i (y_j - \hat{y}_j)^2}{\sum_i (y_j - \bar{y}_j)^2} \quad (1)$$

Results for the R-squared (R^2) metrics: R^2 usually has a range of 0 to 1. Figure 5 shows graph for R^2. All eight approaches yielded a value ranging from -0.5 to 1 for R^2. R^2 values that are negative indicate that the fitted models are worse than the average fitted model. KNN with value 0.9742 is the approach that produced the best R^2 value. When computing the missing value in KNN, the K is set to 4, implying that the value for a missing point is computed using four nearest neighbors. DataWig, on the other side with an R^2 of -0.5311, had the poorest performance. SimpleImputer, Hot Deck, MICE, and Random Forest Regression all received positive results, with values of 0.9744, 1.0, 0.9929, 0.97443, and 0.9745, respectively. Linear Regression and MissForest, on the other hand, calculated negative R^2 values of -0.4356 and -0.0259, respectively.

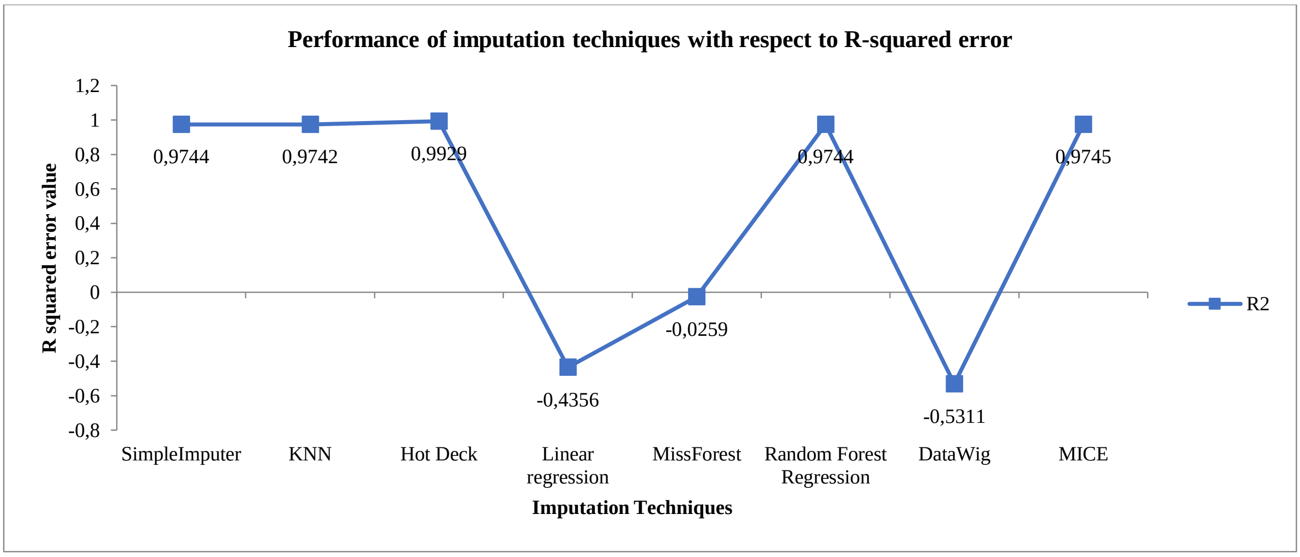


Figure 5: Graphical representation of comparison of imputation techniques with respect to R-squared error.

b. Mean Squared Error: The Mean Squared Error (MSE) is one of the most basic and often used loss functions. To calculate the MSE, take the difference between model's predictions and the ground truth, square it, and average it over the whole dataset. The value of MSE can never be negative because errors are always squared. The amount of samples tested is denoted by N. The advantage with MSE is that it is useful for ensuring that our trained model does not contain any outlier predictions with significant mistakes, as the squaring element of the function gives these errors more weight. The MSE is mathematically defined by the equation 2:

$$MSE = \frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2 \qquad (2)$$

Result for Mean squared Error (MSE) metrics: The Mean Squared Error ranges from 0 to infinity. Figure 6 shows graph for MSE. The value point for MissForest is out of the range when compared to the other points; hence it isn't depicted in this graph. The MSE regression is the most widely used regression for loss functions. Because the real and predicted values are so near, the lower the MSE value, the higher the predicted values accuracy. MissForest (1207.2801) is the strategy that produced the highest MSE while Hot Deck (0.0145) produced the lowest value. MSEs are smaller than 1 for SimpleImputer (0.0514), KNN (0.0529), Random Forest regression (0.0515), and MICE (0.0513) and Linear Regression (1.2888) and DataWig (1.3746) have MSEs more than 1.

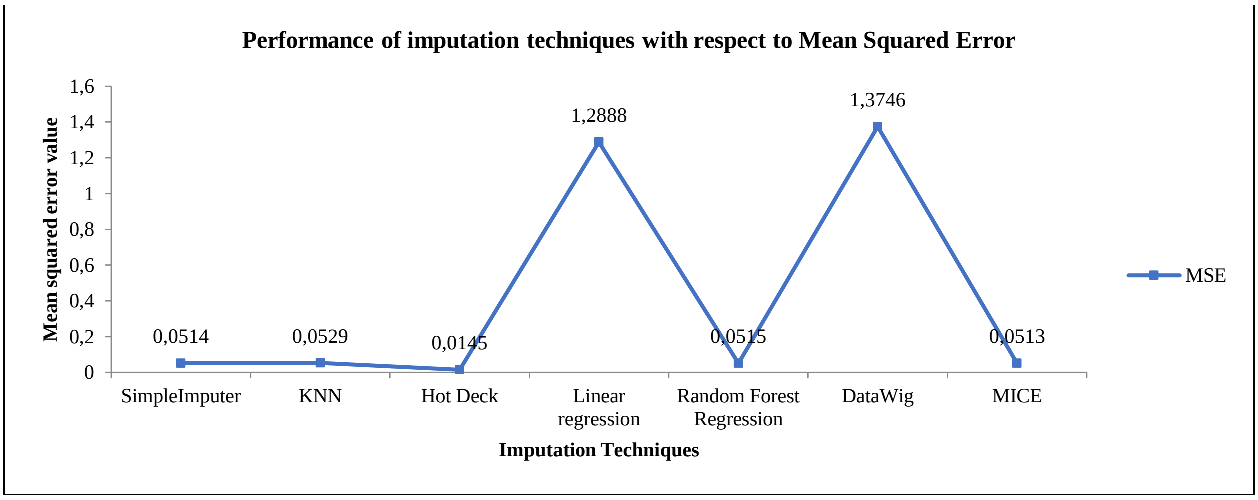


Figure 6: Graphical representation of comparison of imputation techniques with respect to MSE.

c. Mean Absolute Error: The difference between the model's predictions and the ground truth is used while computing the Mean Absolute Error (MAE) and the absolute value is applied to the difference and averaged throughout the entire dataset. The MAE advantage compensates for the MSE disadvantage directly. Because the absolute value is considered, all errors will be weighted on the same linear scale. As a result, unlike the MSE, the loss function will not place an excessive emphasis on outliers and will provide a general and consistent evaluation of how well our model is performing. The MAE is mathematically defined by the equation 3:

$$MAE = \frac{1}{N}\sum_{j=1}^{N}|y_j - \hat{y}_j| \qquad (3)$$

Results for Mean Absolute Error (MAE) metrics: Mean absolute error ranges from 0 to infinity. Figure 7 shows graph for MAE. Initially, the MAE error is calculated in phases. By subtracting the predicting value from the actual value, the prediction error is calculated. Then, for each imputation, the prediction error is calculated and transformed to positive values. It is determined what the mean of all absolute errors is. The best MAE results were achieved by Hot Deck (0.0052), while the poorest MAE results was achieved by MissForest (7.6032). Other techniques produced results ranging from 0 to 1 such as MICE (0.0410), SimpleImputer (0.0411), KNN (0.0245), Linear Regression (1.0319), Random Forest Regression (0.0410) and DataWig (1.0768). The result of measuring the difference between any two continuous variables is generally referred to as Mean Absolute Error.

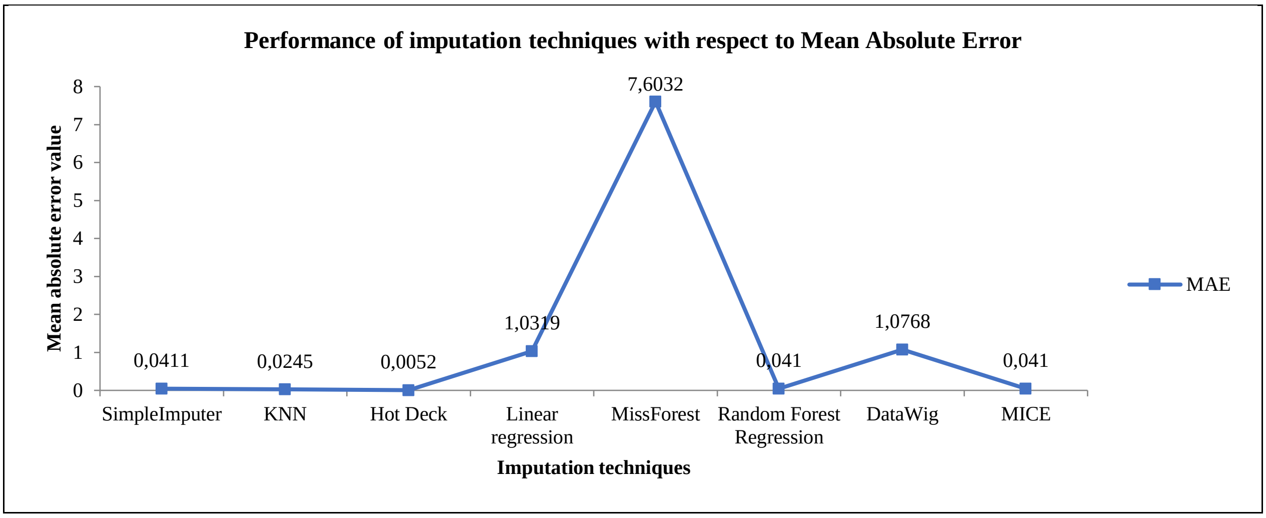


Figure 7: Graphical representation of comparison of imputation techniques with respect to MAE.

As shown in Table 2 Hot Deck imputation technique is the best technique that provides the most promising outcomes and should be considered further, while MissForest produced the worst results. All of the other strategies produced outcomes that might be improved over time by making simple adjustments.

Table 2: Performance comparison of imputation techniques

Techniques	R^2	MSE	MAE
SimpleImputer	0.9744	0.0514	0.0411
KNN	0.9742	0.0529	0.0245
Hot Deck	0.9929	0.0145	0.0052
Linear regression	-0.4356	1.2888	1.0319
MissForest	-0.0259	1207.2801	7.6032
Random Forest Regression	0.9744	0.0515	0.0410
DataWig	-0.5311	1.3746	1.0768
MICE	0.9745	0.0513	0.0410

6 Conclusion

When a value in a dataset goes missing, important information is lost. To avoid this, missing values are imputed. The term "imputing values" refers to the statistical computation of a value for a missing value based on surrounding values or values from the same column. In data analysis, post imputation is significant because it ensures that the dataset is complete and that the findings are computed and arranged accurately. Eight techniques have been explored in this experiment to compute missing values for the Amazon dataset. Only the three columns (Overall, Verified, and Vote) have been utilized to conduct the experiment. Overall column contains missing values, and hence is the most essential column. After imputing the missing values accurately, the outcomes have been evaluated using three evaluation parameters-R^2, MAE and MSE. Hot Deck Imputation technique has surpassed all other techniques in terms of imputation results. The performance metrics for Hot Deck are within the range; however, MissForest’s values are outside the range, making it the lowest performing technique.

References

[1] Afrifa-Yamoah, E. et al. 2020. Missing data imputation of high-resolution temporal climate time series data. *Meteorological Applications*. 27, 1 (2020), 1–18. DOI:https://doi.org/10.1002/met.1873.

[2] Bießmann, F. et al. 2019. DataWig: Missing value imputation for tables. *Journal of Machine Learning Research*. 20, (2019), 1–6.

[3] Burgette, L.F. and Reiter, J.P. 2010. Multiple Imputation for Missing Data via Sequential Regression Trees. *American Journal of Epidemiology*. 172, 9 (Nov. 2010), 1070–1076. DOI:https://doi.org/10.1093/AJE/KWQ260.

[4] Chhabra, G. et al. 2019. A review on missing data value estimation using imputation algorithm. *Journal of Advanced Research in Dynamical and Control Systems*. 11, 7 Special Issue (2019), 312–318.

[5] Cismondi, F. et al. 2013. Missing data in medical databases: Impute, delete or classify? *Artificial Intelligence in Medicine*. 58, 1 (2013), 63–72. DOI:https://doi.org/10.1016/j.artmed.2013.01.003.

[6] Ghazanfar, M.A. and Prugel-Bennett, A. 2013. The advantage of careful imputation sources in sparse data-environment of recommender systems: Generating improved SVD-based recommendations. *Informatica (Slovenia)*. 37, 1 (2013), 61–92.

[7] Graham, J.W. et al. 2003. Methods for Handling Missing Data. *Handbook of Psychology*. (2003).

DOI:https://doi.org/10.1002/0471264385.wei0204.

[8] Heitjan, D.F. and Basu, S. 1996. Distinguishing "missing at random" and "missing completely at random." *American Statistician*. 50, 3 (1996), 207–213. DOI:https://doi.org/10.1080/00031305.1996.10474381.

[9] Jakobsen, J.C. et al. 2017. When and how should multiple imputation be used for handling missing data in randomised clinical trials - A practical guide with flowcharts. *BMC Medical Research Methodology*. 17, 1 (2017), 1–10. DOI:https://doi.org/10.1186/s12874-017-0442-1.

[10] Kaiser, J. 2014. Dealing with Missing Values in Data. *Journal of Systems Integration*. (2014), 42–51. DOI:https://doi.org/10.20470/jsi.v5i1.178.

[11] Kang, H. 2013. The prevention and handling of the missing data. *Korean Journal of Anesthesiology*. 64, 5 (2013), 402. DOI:https://doi.org/10.4097/kjae.2013.64.5.402.

[12] Kim, K.Y. et al. 2004. Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC Bioinformatics*. 5, 1 (Oct. 2004), 1–9. DOI:https://doi.org/10.1186/1471-2105-5-160/FIGURES/3.

[13] Lin, W.C. and Tsai, C.F. 2020. Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*. 53, 2 (2020), 1487–1509. DOI:https://doi.org/10.1007/s10462-019-09709-4.

[14] Little, R.J.A. and Rubin, D.B. 2014. Statistical analysis with missing data. *Statistical Analysis with Missing Data*. (Jan. 2014), 1–381. DOI:https://doi.org/10.1002/9781119013563.

[15] Mandel J, S.P. 2015. A Comparison of Six Methods for Missing Data Imputation. *Journal of Biometrics & Biostatistics*. 06, 01 (2015), 1–6. DOI:https://doi.org/10.4172/2155-6180.1000224.

[16] McAuley, J. et al. 2015. Image-based recommendations on styles and substitutes. *SIGIR 2015 - Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. (2015), 43–52. DOI:https://doi.org/10.1145/2766462.2767755.

[17] Myers, T.A. 2011. Goodbye, Listwise Deletion: Presenting Hot Deck Imputation as an Easy and Effective Tool for Handling Missing Data. *Communication Methods and Measures*. 5, 4 (2011), 297–310. DOI:https://doi.org/10.1080/19312458.2011.624490.

[18] Plaia, A. and Bondì, A.L. 2006. Single imputation method of missing values in environmental pollution data sets. *Atmospheric Environment*. 40, 38 (2006), 7316–7330. DOI:https://doi.org/10.1016/j.atmosenv.2006.06.040.

[19] Ropper, A.H. et al. 2012. Hyperosmolar Therapy for Raised Intracranial Pressure. *New England Journal of Medicine*. 367, 26 (2012), 2554–2557. DOI:https://doi.org/10.1056/nejmc1212351.

[20] Schuetz, C.G. 2008. Using neuroimaging to predict relapse to smoking: role of possible moderators and mediators. *International journal of methods in psychiatric research*. 17 Suppl 1, 1 (2008), S78–S82. DOI:https://doi.org/10.1002/mpr.

[21] Sinharay, S. et al. 2001. The use of multiple imputation for the analysis of missing data. *Psychological Methods*. 6, 3 (2001), 317–329. DOI:https://doi.org/10.1037/1082-989x.6.4.317.

[22] Stekhoven, D.J. and Bühlmann, P. 2012. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 28, 1 (Jan. 2012), 112–118. DOI:https://doi.org/10.1093/BIOINFORMATICS/BTR597.

[23] Tan, Y. et al. 2018. Probability matrix decomposition based collaborative filtering recommendation algorithm. *Informatica (Slovenia)*. 42, 2 (2018), 265–271.

[24] Troyanskaya, O. et al. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics*. 17, 6 (Jun. 2001), 520–525. DOI:https://doi.org/10.1093/BIOINFORMATICS/17.6.520.

[25] Zhang, Z. 2016. Missing data imputation: Focusing on single imputation. *Annals of Translational Medicine*. 4, 1 (2016). DOI:https://doi.org/10.3978/j.issn.2305-5839.2015.12.38.

Computational Analysis of Uplink NOMA and OMA for 5G Applications: An Optimized Network

Shelesh Krishna Saraswat, Vinay Kumar Deolia, Aasheesh Shukla
[1]Department of Electronics and Communication Engineering, GLA University, Mathura (281406), UP, India.

E-mail: shelesh.saraswat@gla.ac.in, vinaykumar.deolia@gla.ac.in, aasheesh.shukla@gla.ac.in

Keywords: Non-orthogonal Multiple Access (NOMA), Multiple orthogonal Access (MOA), Jain index

In this paper, the non-orthogonal multiple access (NOMA) schemes are compared with the multiple orthogonal access (OMA) schemes on the basis of the resource allocation validity of uplinks. By reflecting the involvement of a measure of each user's data on the system's total amount, we analyze the main reasons why NOMA provides justice service distribution over OMA on unequal channels. Moreover, the Jain index is observed and proposed to quantify the irregularity of numerous user channels, according to the metric for the Jain index based on the Jain index. More importantly, the proposed metric establishes the criteria for choosing between NOMA and OMA to share resources correctly. Based on this debate, we offer a program that combines NOMA and OMA to increase user integrity. Imitation effects substantiate the exactness of the proposed matrix and display improvement of the accuracy of the showcased NOMA-OMA mixture system as compared to standard OMA as well as NOMA systems. The Biggest technology development in the next years is the Internet of Things, which promises omnipresent connectivity of everything everywhere. it's anticipated that over 25 billion gadgets will be linked to cellular networks. Various challenges are faced by the wireless networks of Fifth generation (5G) , the main challenge discussed in this paper regarding channel fetching schemes. For massive connectivity so that we can increase data rate and save bandwidth also.

Povzetek: V tem članku so primerjane sheme neortogonalnega (NOMA) in ortogonalnega dostopa (OMA) glede na veljavnost dodeljevanja virov pri povezavah.

1 Introduction& literature studies

For 5th generation (5G) wireless networks, the non-orthogonal multiple access (NOMA) schemewas identified as a favourablereflection of a multi-access system welcoming extra user and enhancing efficiency in spectral manner [1] - [5]. In the first version of this type of technique, the multiple user's superposition transmission system (MUST), wasshowed for a 3-year partnership project advanced evolution networks i.e., 3GPP-LTE-A [6]. The basic impression of this technique is to take advantage in the field of power in order to cultivate two things: first thing should be used in multiple user multiplexing and the second thing to employ user intervention (SIC) for persistent disturbance (IUI) cancellation. In contrast to standard schemes for orthogonal multiple access (OMA) [7],[8], NOMA enables multiple transmissions simultaneously Superposition coding with varying power levels allows users to have the same degree of freedom (DOF). Meanwhile, through exploitation, advanced signal processing methods, such as SIC, can compensate for the received power differential to obtain the appropriate signals to the recipient. Compared to traditional OMA systems, NOMA has been shown to significantly increase the system's spectral efficiency [9] - [11]. Consequently, NOMA I can support large connections, minimize latency in communication, and increase efficiency of the system by spectral means. Most current activities reduce NOMA programs [9] - [12].

On the other hand, NOMA is found in abundance in the uplink communication, where the electric waves are naturally placed at the top of the various forces received at the reception base station (BS). Aside from that, SIC recordings are usually more economical for BSs than they are for mobile customers. In [13], the authors compare NOMA with OMA regarding spectral energy point efficiency in the uplink. Lately, the researchers of [14], [15] have devised a resource allocation-based distribution technique for most instances (ML) recipients in BS. On the other side, one of the essential aspects of NOMA is that it ensures equity in resource distribution. Unlike OMA systems, which allow customers with bad channel conditions to be halted on service, NOMAlets users with varied channel settings to be served concurrently. In Ku [16], the NOMA uplink system presented a power allocation mechanism to provide users with max-min justice. Ku [17] looked into editing with a system of non-orthogonal repetitious users who were inaccurate. In

[18], [19], investigative power distribution was investigated on one side, and many NOMA sticks below systems, respectively. Even though, the fairness was considered by the resource allocation-based distribution technique [16] - [21], it was still uncertainabout NOMA to offer more resource allocation as compared to OMA.

This paper wants to see how NOMA and OMA compare service accuracy uplink sharing. A selection condition is proposed when delivering status information for the current channel to determine if NOMA or OMA should be employed. By presenting the sacrifice of perfection individual user data rate system rating, we state why NOMA is fair to them service distribution than OMA on unequal channels. In addition, we offer a measure for the accurate indication of a closed-form for deciding when NOMA is superior to OMA for two users of the NOMA1 system. Plus, a simple hybrid NOMA-OMA program that selects NOMA flexibly once The OMA in terms of the proposed metrics are proposed to continue improving user integrity. The numerical results are displayed to verify our proposed matrix's accuracy and improve the merits of the proposed NOMAOMA integrated system. Some negatives are such as Overload and Preamble Collision Problems, Excessive Overhead , more QoS Requirements, Power radiations.The following is the order in which the paper is organized. The NOMA system uplink system and the chat NOMA and OMA capacity areas were presented in section II. The reason why NOMA is just more efficient than OMA is analyzed in section III. Alternatively, the metric indicator for the accuracy of the closed-form and the hybrid NOMA-OMA is a program that has been recommended. In Section IV, simulation effects are introduced and investigated. Finally, this paper is concluded in section V.

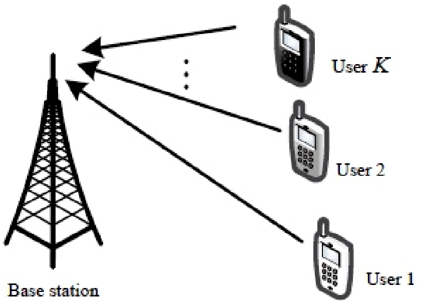


Figure 1: An uplink of NOMA model with a base station and K users.

The symbols used in this paper are as follows. Circular Gaussian distribution of the same complex with the mean μ and variance σ^2 is defined as $\mathcal{CN}$ (μ, σ^2); $\sim$ it must be "distributed as "; $\mathbb{C}$ stands for a collection of all the complex numbers; $|\cdot|$ describes the total amount of the complex scalar; Pr $\{\cdot\}$ mean chances of random occurrence. Smart indoor communications, remote area communication, smart outdoor communication for smart city Auto-pilot UAVs, Self-driving Electrical Vehicles, Fast Regional Trains are some relevant examples.

2 Model of noma and oma system

The NOMA model for uplink is introduced in this part along with the NOMA, OMA power zones.

A. System Model

As indicated in Figure, we're putting the NOMA uplink technology to the test with single-antenna BS and K users.All K users send within one network companywith the same transmission power (P_0). For the NOMAsystem, K no. of users is repeated in the identical network company with dissimilar power levels are not accepted. In contrast, in the OMA system, K users use a network company that uses a time-sharing strategy [22]. The signal received on BS is given by the NOMA system.

$$y = \sum_{k=1}^{K} \sqrt{p_k} h_k s_k + v \qquad (1)$$

where $h_k \in \mathbb{C}$is defined as the channel coefficientsbetween BS and the user, and $k = \{1, \dots \dots, K\}$ is the channel coefficient between BS and the user, s_k defines a modified symbol to user k, p_k means user transfer k, and $v \sim \mathcal{CN}$ (0, σ^2) means an additional white Gaussian sound (AWGN) in BS and σ^2 sound power. Without losing common sense, we think $|h_1|^2 \leq |h_2|^2 \leq \cdots |h_K|^2$.

B. Region of Power

The OMA system is well-known for optimal DOF allocation and multimedia NOMA application. As demonstrated in Figure 2, the power supply can reach the same quantity of system uplink transmission [22], [23]. Here is the full-service offer for either NOMA and OMA techniques.

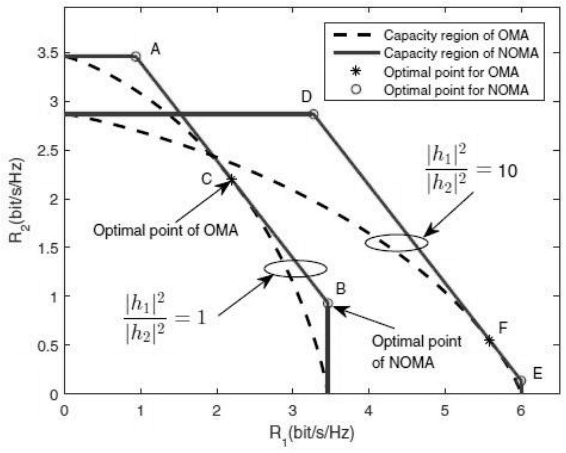


Figure 2: The capacity region of NOMA and OMA for realization of a single channel with one BS plus two users.

The capacity region having NOMA as well OMA for a single channel realization with a BS plus two users illustrated in Fig. 2. The two users have transmitted power of $P_0 = 20\,dB$. When the curve of $\left|\frac{h_1}{h_2}\right|^2 = 1$, we have $\frac{|h_1|^2}{\sigma^2} = \frac{|h_2|^2}{\sigma^2} = 20\,dB$. For the curve of $\left|\frac{h_1}{h_2}\right|^2 = 10$, we have $\frac{|h_1|^2}{\sigma^2} = 18\,dB$, and $\frac{|h_2|^2}{\sigma^2} = 28\,dB$.

$$\alpha_k = \frac{|h_k|^2}{\sum_{i=1}^{K} |h_i|^2};\ \forall k \qquad (2)$$

It is noted that α_k can be translated to the as a normal channel benefit k. In other terms, a fair share of DOF through the OMA program to share network company with duration in proportion to their normal channel benefits, and depends on the distribution of time variables according to immediate channel fulfillment. We recognize that it is correct The DOF distribution is available to all users who submit via their transmission capacity is P_0 as there is no IUI in its OMA system.

Furthermore, by considering $p_k = P_0, \forall k$, and executing SIC in BS [22], [24], an effective NOMA power distribution corner points, namely point A, B, D, and E in the middle Fig. 2, may be found. A time-sharing approach can be used to gain any pair of scales in line segments between existing locations. Even though OMA's regional capacity is lower than NOMA's, Fig. 2 indicates that NOMA regularly beats OMA on the basis of spectral efficiency and its user bias, thanks to its time-sharing mechanism. We should emphasize that NOMA can only obtain corner points in the power field without a time-sharing method, resulting in less fairness than OMA in instances.

This paper showcase about the legitimacy of NOMA usersystems with absence of time-sharing and the OMA system with flexible DOF distribution. Both methods receive the similar system sum-rate but lead to unlike users' justice. With understanding, in Fig. 2, of a channel equal to $\left|\frac{h_1}{h_2}\right|^2 = 1$, OMA in area C is better than NOMA as both the users have the similar amount of specific data. However, with channel of asymmetric having$\left|\frac{h_1}{h_2}\right|^2 = 10$, It should be observed that OR in point D is preferable to OMA in the right place F. Consequently, it is fascinating to present explanations for justice development of NOMA on unequal channels and availability of a quantifiable fairness indicator to determine the superiority of NOMA overOMA.

3 Fairness comparison between noma and oma

In this section of the paper, the Jain justice accepted index has been introduced [25] for assessing resource-based allocation goodness. Then, overall rating towards contribution of each and every user data rate in the systems have been presented. The main reasons why NOMA is less biased than OMA. The closed version of the justice index in the NOMA system for two users is based on the Jain index [25] to govern if you are utilizing either NOMA or OMA in combination of users in a single network firm. In addition, a proposed NOMA or hybrid system using NOMA or OMA is proposed flexibly based on the showcased matrix. Because of NOMA technique. Internet speed of communication could be better which can increase the visibility of E-Commerce.

A. Jain's Fairness Index

In this study, Jain's index is used [25] to quantify fairness in the subsequent scenarios.

$$J = \frac{\left(\sum_{k=1}^{K} R_k\right)^2}{K \sum_{k=1}^{K} (R_k)^2} \qquad (3)$$

WhereR_k refers to each user level k. Note that $\frac{1}{K} \le J \le 1$. A system with a greater Jain index is very good and reaches the extreme when each and very users receive the same amount of specific data.

B. Analyzing Righteousness

For the full-service offer, both NOMA as well as OMA programs, deliberated in part II-B, are readily availablethe total amount and data levels for each of the two schemes as follows:

$$R_{sum}^{NOMA} = R_{sum}^{OMA} = \sum_{i=1}^{K} R_k^{NOMA} = \sum_{i=1}^{K} R_k^{OMA} = \log_2\left(1 + \frac{P_0}{\sigma^2}\sum_{i=1}^{K}|h_i|^2\right) \tag{4}$$

$$R_k^{NOMA} = \log_2\left(1 + \frac{P_0|h_k|^2}{P_0\sum_{i=1}^{k-1}|h_i|^2 + \sigma^2}\right) \tag{5}$$

$$R_k^{OMA} = \alpha_k R_{sum}^{OMA} \tag{6}$$

Wherethe entire program rating of NOMA and OMA schemes with suitable resource allocation is referred to as R_{sum}^{NOMA} and R_{sum}^{OMA},R_k^{NOMA} and R_k^{OMA} in NOMA and OMA applications, respectively, refer to a measure of user data. We first define the collection of normal channel gain in the NOMA program, such as$\emptyset_k = \sum_{i=1}^{k}\alpha_i\,, k = \{1, \dots, K\}$,$\emptyset_0 = 0$, then rewrite the user-accessible number k as

$$R_k^{NOMA} = \log_2\left(1 + \frac{P_0\emptyset_k}{\sigma^2}\sum_{i=1}^{K}|h_i|^2\right) - \log_2\left(1 + \frac{P_0\emptyset_{k-1}}{\sigma^2}\sum_{i=1}^{K}|h_i|^2\right) \tag{7}$$

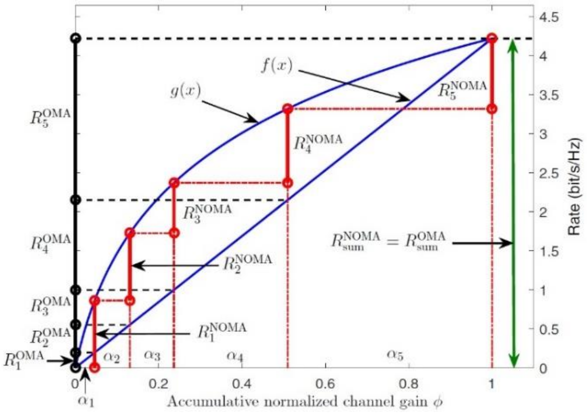


Figure 3: Diagram of the total level of the system against accumulative normalized NOMA, OMA channel benefits with $K = 5$users. The green double. arrow shows total NOMA program and OMA program values . The NOMA and OMA program's ratings are depicted by red and black line segments, respectively.

The first term (7) denotes the system's total value of users, and the second term refers to a system having$k - 1$number of users. To put it another way, the role of user k to overall system rating is determined by the logarithm function difference concerning $\emptyset_k$ and $\emptyset_{k-1}$. We explain the logarithm function as follows for the sake of simplicity and generality.

$$g(x) = \log_2(1 + \Gamma x);\;\; 0 \le x \le 1 \tag{8}$$

with

$$\Gamma = \frac{P_0}{\sigma^2}\sum_{i=1}^{K}|h_i|^2\ ;$$

$$R_k^{NOMA} = g(\emptyset_k) - g(\emptyset_{k-1}) \tag{9}$$

Furthermore, OMA program, it is seenfrom (6) to R_k^{OMA} has a line in association with $R_{sum}^{OMA}$and the slope w.r.t. the total amount of the system is determined by normal channel gain $\alpha_k = \emptyset_k - \emptyset_{k-1}$. Similarly, the difference between the line functions $\emptyset_k$,$\emptyset_{k-1}$ determines the user contribution k in overall system rating, where

$$f(x) = \log_2(1 + \Gamma)x\,;\; 0 \le x \le 1\,; \text{ and}$$

$$R_k^{OMA} = f(\emptyset_k) - f(\emptyset_{k-1}) \tag{10}$$

Fig. 3 depicts the linear as well as logarithmic rise in model data rate in OMA and NOMA with $K = 5$ uplink users as a function of accumulated channel profits. It is noteworthy that NOMA and OMA programs have four the total amount of the same system but given a different date individual user number. In particular, the NOMA system gains a better service share than the OMA system because all users are assigned the same person's prices. The logarithmic map of $g(\emptyset_k)$ concerning Accumulated channel gain $\emptyset_k$, in reality, helps the fairness of service sharing in NOMA. The first and second outputs, respectively, of which $g(\emptyset_k)$ increases and decreases concerning $\emptyset_k$. Large uza normal channel gain kuhamba, slow $g(\emptyset_k)$ increase by $\emptyset_k$, When compared to the OMA method, this results in a modest sum per individual. On the other hand, a small normal gain of the channel α_k can lead to an increase increasing level of $g(\emptyset_k)$ by $\emptyset_k$, the higher a person the rate is attainedrelated to that of the OMA system. Because for example, it is considered a weak user and a very strong user with standard channel gain of α_1 and α_K, respectively, R_1^{NOMA} suggested logarithm function $g(x)$ in comparison to R_1^{NOMA}, when compared to R_K^{OMA}, R_K^{NOMA} is lower.

Note 1: It's worth noting that OMA line mapping is more straightforward than the NOMA program for equal channels. However, the chances are that all users are the same. The benefits of the channel are very small, particularly for a program with a huge user base

C. Metric for Fairness Indicator

In reality, most NOMA programs believe at least two users repeat with the similar DOF [11, 12, 26], which can minimize both computational difficulty and recipient coding latency. As a result, in this portion, we concentrate on a fair comparison of NOMA as well as OMA with $K = 2$. We'd like to construct simple criteria for determining whether NOMA is significantly better than OMA for two users, for which it is crucial for improving user planning in a system having multiple DOFs and users. The following theorem proposes righteousness as a metric index. Theory 1: If fewer users are provided with channel $|h_2|^2 \geq |h_1|^2$, the NOMA system is really fair to a strong logic of Jain's right to righteousness only if.

$$\frac{|h_1|^2}{|h_2|^2} \leq \frac{\beta}{1-\beta} \tag{11}$$

where

$\beta = \dfrac{W\left(\frac{(1+\Gamma)^{1+\frac{1}{\Gamma}}\log(1+\Gamma)}{\Gamma}\right)}{\log(1+\Gamma)} - \frac{1}{\Gamma}$ and $W(x)$: the Lambert W function. For high SNR regime, i.e., $\Gamma \to \infty$, We have an approximation of β as with a high SNR.

$$\tilde{\beta} = \frac{W(\log(1+\Gamma))}{\log(1+\Gamma)} \tag{12}$$

Proof: Because the sum of the NOMA and OMA schemes is the same, we must compare the total square of individual ratings (SSR), that is, $SSR = \sum_{k=1}^{2}(R_k)^2$, the denominator value of the NOMA scheme (3). A system with a modest SSR can be skewed in Jain's way. Through the OMA program, Sine

$$SSR_{OMA} = (\log_2(1+\Gamma))^2(\alpha_1^2 + \alpha_2^2)$$
$$= (\log_2(1+\Gamma))^2(1 + 2\alpha_1^2 - 2\alpha_1) \tag{13}$$

where $0 \leq \alpha_1 \leq 0.5$ since we assume $|h_1|^2 \leq |h_2|^2$. The SSR_{NOMA} can be used by, in the NOMA scheme-

$$SSR_{NOMA} = (\log_2(1+\Gamma\alpha_1))^2 + (\log_2(1+\Gamma) - \log_2(1+\Gamma\alpha_1))^2$$
$$= (\log_2(1+\Gamma))^2 + 2\log_2(1+\Gamma\alpha_1))^2 - 2\log_2(1+\Gamma)\log_2(1+\Gamma\alpha_1) \tag{14}$$

It's worth noting that the smaller SSROMA solution = SSRNOMA has $\alpha_1 = 0$, which corresponds to a single user situation. Moreover, in $\alpha_1 = 0.5$, that is,$|h_1|^2 = |h_2|^2$, we have$SSR_{OMA} < SSR_{NOMA}$ as observed in the volume region of Figure 2. In addition, SSR_{OMA} is a monotonic entity decrease between $0 \leq \alpha_1 \leq 0.5$, while SSR_{NOMA} is a monotonic degradation function of α_1 within $0 \leq \alpha_1 \leq \frac{\sqrt{1+\Gamma}-1}{\Gamma}$and increases by α_1 within $\frac{\sqrt{1+\Gamma}-1}{\Gamma} \leq \alpha_1 \leq 0.5$. And, from Figure 2, we can witness that $SSR_{OMA} > SSR_{OMA}$ of small positive negligence α_1. Therefore, there is a unique combination of SSR_{OMA} and SSR_{NOMA} at $\alpha_1 = \beta$ in the range $0 \leq \alpha_1 \leq 0.5$. Before intersection, i.e., $0 < \alpha_1 < \beta$, NOMA is best, after that at a crossroads, i.e., $\beta < \alpha_1 < 0.5$, OMA is very good. Solving $SSR_{OMA} = SSR_{NOMA}$ within $0 \leq \alpha_1 \leq 0.5$, we find

$$\beta = \frac{W\left(\frac{(1+\Gamma)^{1+\frac{1}{\Gamma}}\log(1+\Gamma)}{\Gamma}\right)}{\log(1+\Gamma)} - \frac{1}{\Gamma}$$

Moreover, at $\alpha_1 \leq \beta$, we have $\frac{|h_1|^2}{|h_2|^2} \leq \frac{\beta}{1-\beta}$, i.e., completes the proof of adequacy of the proposed fairness metric indicator. As needed, the region between $0 < \alpha_1 < 0.5$ where $SSR_{OMA} > SSR_{NOMA}$ is $0 < \alpha_1 < \beta$ is different, which is the sole region between $0 < \alpha_1 < 0.5$. In other sense, NOMA is only useful if $0 < \alpha_1 < \beta$ is true, demonstrating the requirement for this suggested metric.

Note 2: It should be noted that the proposed accuracy index is exclusively dependent on the parameter Γ described in (9). As an outcome, the statistic is focused on the immediate earnings of the channel. In comparison to the Jains, our proposed metrics, which comprise OMA and NOMA, provide greater insight. Especially at the top SNR limitations (12), we can see that $\tilde{\beta}$ decreases by a significant increase in transmission power from Lambert. The W function in number rises slightly higher than that of the W denominator. Therefore, the chances of NOMA being the best will fall when the top post is promoted power, which will be guaranteed in imitation.

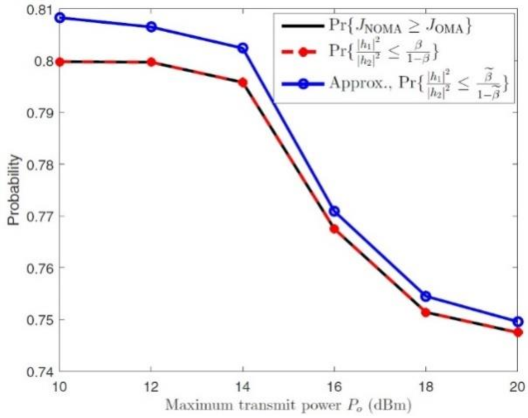


Figure 4: The probability of NOMA being fairer than OMA versus the maximum transmit power, *P*0. D. Hybrid OR OMA Scheme.

4 Simulation results

We use simulation in this section to ensure that the suggested matrix and test hybrid OR OMA method are both effective. One cell with BS available center with cell radius 400 m is taken into consideration. There are$N_F = 128$ sub-system carriers,$2N_F$ users are arbitrarilypaired across all subcarriers. In a cell, all $2N_F$ users are dispersed at random and uniform manner. Under BS, we set the volume of each carrier to $\sigma^2 = -90$dBm. 3GPP path loss model in the form of large urban cellsaccepted into our estimates [27]. Figure 4 shows the potential for OR greater is better than OMA compared to high transmission capacity, P_0. It is worth noting that Pr $\left\{\frac{|h_1|^2}{|h_2|^2} \leq \frac{\beta}{1-\beta}\right\}$ fits well with $Pr\,\{J_{NOMA} \geq J_{OMA}\}$. Simply put, our proposed goodness meter index can guess if NOMA is quite accurate as compared to OMA. Also, with increased SNR,$\tilde{\beta}$ values in equation (12) wherePr $\left\{\frac{|h_1|^2}{|h_2|^2} \leq \frac{\tilde{\beta}}{1-\tilde{\beta}}\right\}$strictly related to imitation effects. Furthermore, according to the Jain index, the NOMA scheme has a higher likelihood of better justice (0.75~ 0.8) than the OMA model. This is because of the fact that the chances of the channels are asymmetric, much largest than the equivalent channels. Furthermore, high transfer power reduces the odds of NOMA being biased, as stated in Remark 2: When contrasted to a high-transfer-power system, this is owing to the NOMA system's limited interference. They are powerless because the user (with high acquired power) will be subjected to a significant level of distraction, whilst SIC will not influence the weak user (poor power is not accepted) recording. As a result, in high transfer capacity, the weak user can

Theorem 1 presents a metaphor for the accuracy index that simplifies assessing if NOMA is superior to OMA and will serve as a condition of the user's schedule design systems that have multiple network companies that provide multiple users, In particular, with the wrong user planning strategy, we suggest a flexible mix scheme that determines each pair users in each network company in the selection or OMA system or a NOMA program to improve user integrity. Instead of using the NOMA program or OMA system in all areas with fewer carriers, this NOMA-OMA mixed program can improve I very user compatibility. It should be noted that it can be continuously enhanced if fairness is developed in conjunction with user editing. Future efforts will be considered.

Achieve a considerably greater data rate than a strong user, which is possible the result of slightly better service delivery than OMA. However, NOMA is still better than OMA possibly about 0.75 in the maximum transmission system. Figure 5 shows the potential for congestion work (PDF) of the user rating of the company's network system with random pairing. The NOMA, OMA as well as the hybrid NOMA-OMA systems (which is proposed in this paper) are all comparable multiple access strategies. Apparently, the average data distribution of each NOMA model is higher and more focused as compared to the OMA model, i.e. The NOMA system provides a more equitable resource-based allocation than OMA system. In general, the distribution of each level of the combined NOMA-OMA hybrid system is quiteintensive than its counterpart i.e., NOMA system. In actuality, the combination we've proposed is a bit of a mishmash. Based on the metric index of accuracy, you can switch between NOMA, OMA as well as the NOMA-OMA program. It can make better use of the channel's benefits connection. The values$J_{NOMA} = 0.76, J_{OMA} = 0.62$ is taken into consideration, whereas the value of Jain index is considered as $J_{Hybrid} = 0.91$ for NOMA-OMA hybrid model. The above-mentioned values are the results of three multiple access strategies.

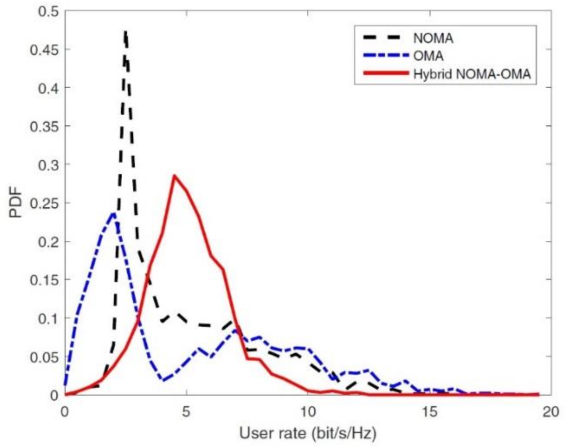


Figure 5: The user rate's PDF for NOMA, OMA, as well as Hybrid NOMA-OMA model

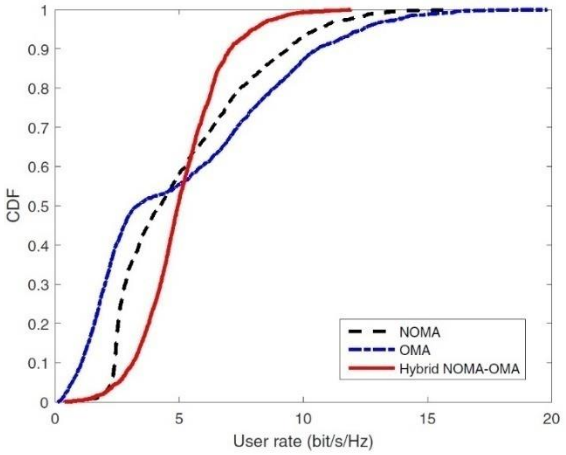


Figure 6: The CDF of the NOMA scheme, the OMA scheme, and the hybrid NOMA-OMA scheme

Furthermore, the user level increasing distribution function (CDF) is approx. It was fascinating in operation, as illustrated in Figure 6. Compared to the NOMA system, the 10-percentile user level, which has a tenuous connection to fairness and user experience, increases by around 1 bit/Hz/s. This suggests that our suggested NOMA-OMA hybrid method can provide significantly better performance for low-level users while also improving user data quality. With the aid of cmos, we may employ this method for radiation losses. Different modulation techniques may enable us to develop a more effective E-Commerce solution in the future.

5 Conclusion

The resource distribution inequalities between NOMA as well as OMA programs into the uplinks are examined in this study. By inserting the characters of the influence of each user's data rate to the total system rating, the basic reason why NOMA is more suited than OMA in irregular multiple user channels was studied. A logarithmic map within the typical channel benefits and estimates of each data that utilizes the channel gains asymmetry is utilized to increase user integrity in NOMA system. On the basis of this remark, we have raised the value fairness indicator metric for NOMA systems for two users determines whether NOMA provides a more equitable service distribution than OMA. Furthermore, we offered a NOMA-OMA mix, which flexibly selects NOMA and OMA for user development goodness based on the provided technique. When NOMA is less biased than this OMA, our recommendation metric can reliably detect it. According to numerical results. Otherwise, a proposed mixed NOMA-OMA system can considerably increase user integrity compared to traditional NOMA-OMA schemes. The enormous economic advantages of mobile commerce are clear when 5G and A IoT technology are combined. The rapid rise of mobile commerce made possible by 5G's high speed, vast capacity, and low latency

References

[1] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-Lin, and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," IEEE Commun. Mag., vol. 53, no. 9, pp. 74–81, Sep. 2015.

[2] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, C. L. I, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," IEEE Commun. Mag., vol. 55, no. 2, pp. 185–191, Feb. 2017.

[3] Z. Wei, J. Yuan, D. W. K. Ng, M. Elkashlan, and Z. Ding, "A survey of downlink non-orthogonal multiple access for 5G wireless communication networks," ZTE Communications, vol. 14, no. 4, pp. 17–25, Oct. 2016.

[4] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, "Non-orthogonal multiple access in multi-cell networks: Theory, performance, and practical challenges," arXiv preprint arXiv:1611.01607, 2016.

[5] V. W. Wong, R. Schober, D. W. K. Ng, and L.-C. Wang, Key Technologies for 5G Wireless Systems.Cambridge University Press, 2017.

[6] "Study on downlink multiuser supersition transmission (MUST) for LTE (Release 13)," 3GPP TR 36.859, Tech. Rep., Dec. 2015.

[7] D. W. K. Ng and R. Schober, "Cross-layer scheduling for OFDMA amplify-and-forward relay networks," IEEE Trans. Veh. Technol., vol. 59, no. 3, pp. 1443–1458, Mar. 2010.

[8] D. W. K. Ng, E. S. Lo, and R. Schober, "Energy-efficient resource allocation in OFDMA systems with large numbers of base station antennas," IEEE Trans. Wireless Commun., vol. 11, no. 9, pp. 3292– 3304, Sep. 2012.

[9] Z. Ding, Z. Yang, P. Fan, and H. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," IEEE Signal Process. Lett., vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

[10] Z. Yang, Z. Ding, P. Fan, and G. K. Karagiannidis, "On the performance of non-orthogonal multiple access systems with partial channel information," IEEE Trans. Commun., vol. 64, no. 2, pp. 654–667, Feb. 2016.

[11] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," IEEE Trans. Commun., 2017, accepted for publication.

[12] Z. Wei, D. W. K. Ng, and J. Yuan, "Power-efficient resource allocation for MC-NOMA with statistical channel state information," Proc. IEEE Global Commun. Conf., pp. 1–1, Dec. 2016.

[13] P. Wang, J. Xiao, and L. P, "Comparison of orthogonal and non-orthogonal approaches to future wireless cellular systems," IEEE Veh. Technol. Mag., vol. 1, no. 3, pp. 4–11, Sep. 2006.

[14] M. Al-Imari, P. Xiao, M. A. Imran, and R. Tafazolli, "Uplink non-orthogonal multiple access for 5G wireless networks," in Proc. IEEE Intern.Sympos. on Wireless Commun.Systems, Aug. 2014, pp. 781– 785.

[15] M. Al-Imari, P. Xiao, and M. A. Imran, "Receiver and resource allocation optimization for uplink NOMA in 5G wireless networks," in Proc. IEEE Intern.Sympos. on Wireless Commun. Systems, Aug.2015, pp. 151–155.

[16] N. Zhang, J. Wang, G. Kang, and Y. Liu, "Uplink non-orthogonal multiple access in 5G systems," IEEE Commun.Lett., vol. 20, no. 3, pp. 458–461, Mar. 2016.

[17] T. Takeda and K. Higuchi, "Enhanced user fairness using non-orthogonal access with SIC in cellular uplink," in Proc. IEEE Veh.Techn. Conf., Sep. 2011, pp. 1–5.

[18] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," IEEE Signal Process. Lett., vol. 22, no. 10, pp. 1647–1651, Oct. 2015.

[19] Y. Liu, M. Elkashlan, Z. Ding, and G. K. Karagiannidis, "Fairness of user clustering in MIMO non-orthogonal multiple access systems," IEEE Commun.Lett., vol. 20, no. 7, pp. 1465–1468, Jul. 2016.

[20] P. D. Diamantoulakis, K. N. Pappi, Z. Ding, and G. K. Karagiannidis, "Wireless-powered communications with non-orthogonal multiple access," IEEE Trans. Wireless Commun., vol. 15, no. 12, pp. 8422–8436, Dec. 2016. [21] Y. Yu, H. Chen, Y. Li, Z. Ding, and B. Vucetic, "Antenna selection for MIMO-NOMA networks," arXiv preprint arXiv:1609.07978, 2016.

[22] D. Tse and P. Viswanath, Fundamentals of wireless communication. Cambridge University Press, 2005.

[23] M. Vaezi and H. V. Poor, "Simplified Han-Kobayashi region for onesided and mixed gaussian interference channels," in Proc. IEEE Intern.Commun. Conf., May 2016, pp. 1–6.

[24] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (NOMA) systems," IEEE Access, vol. 4, pp. 6325–6343, Aug. 2016.

[25] T. Lan, D. Kao, M. Chiang, and A. Sabharwal, "An axiomatic theory of fairness in network resource allocation," in Proceedings IEEE INFOCOM, Mar. 2010, pp. 1–9.

[26] Z. Ding, P. Fan, and H. V. Poor, “Impact of user pairing on 5G non-orthogonal multiple-access downlink transmissions,” IEEE Trans. Veh. Technol., vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

[27] “Evolved universal terrestrial radio access: Further advancements for E-UTRA physical layer aspects,” 3GPP TR 36.814, Tech. Rep., 2010.

[28] Mr. Ather Parvez Abdul Khalil. (2012). Healthcare System through Wireless Body Area Networks (WBAN) using Telosb Motes. International Journal of New Practices in management and Engineering, 1(02), 01 – 07.

[29] Mr. Dharmesh Dhabliya, Prof. Ojaswini Ghodkande. (2016). Prevention of Emulation Attack in Cognitive Radio Networks Using Integrated Authentication. International Journal of New Practices in Management and Engineering, 5(04), 06 - 11.

[30] Dr. S.A. Sivakumar. (2019). Hybrid Design and RF Planning for 4G networks using Cell Prioritization Scheme. International Journal of New Practices in Management and Engineering, 8(02), 08 - 15.

Designing A Permissioned Blockchain Network for the Insurance Claim Process Using Hyperledger Fabric and Composer

Archana Hombalimath[1*], Neha Mangla[2], Arun Balodi[2]
[1]CMR Institute of Technology, Bangalore, Research Scholar Atria Institute of Technology, Bangalore
[2]Atria Institute of Technology, Bangalore
E-mail: archana.h@cmrit.ac.in, neha.mangla@atria.edu, drbalodi@gmail.com
[*]Corresponding author

Keywords: blockchain (bc), hyperledger fabric, hyperledger composer, vehicle insurance

Received: May 6, 2022

This research aims to examine the benefits of blockchain technology (BCT) in the vehicle insurance process. The article addresses a number of benefits offered by BCT, including automating the identity verification and doing away with the need for numerous parties to manually certify the legitimacy of transactions. India's financial and business networks can be anticipated to change the accounting process to a different extreme with the introduction of BCT. Unfortunately, they are having some difficulties implementing and acclimating to this modern technology. Here is the solution, blockchain technology may help us solve the existing problem. Insurance firms and car owners can benefit from blockchain technology since it can effectively open up communication channels, encourage industry integration, and improve insurance provider's ability to access record. Using BCT, banking operations will be more efficient, quicker, and less expensive because to the removal of middlemen. Decentralization, transparency, and secure transactions will be the main advantages.

Povzetek: Raziskava preučuje prednosti tehnologije blockchain v procesu avtomobilskega zavarovanja: avtomatizacijo, transparentnost in varnost transakcij.

1 Introduction

1.1 Background

According to the IAIS (International Association of Insurance Supervisors), 20 to 30 percent are fraudulent insurance claims. Clients who lack the ethics and legal knowledge are the one who do the insurance fraud. The policyholder's information is not to be shared. Even if an insurance firm blacklists some policyholders, this does not prevent other insurers from doing business with that insurer. Lack of data transparency and asymmetric information between client and insurance company, the policyholder go for contract.

The existing insurance sector, on the other hand, has convoluted underwriting processes, increased price of underwriting time. The insurance business is not well-monitored. As a consequence, analyzing the risks of insurance applicants while avoiding moral hazard is impossible. It's also unable of dealing effectively with the current problem of criminals exploiting blockchain to launder money.

In the field of automobile insurance, there is a cyclical link between insurers, police, repair shops, and insurance providers. Every connection has challenges with low efficiency and complicated services. Insurance expenses are expensive for insurance service providers, particularly administration costs. Contract signing and management, database maintenance, payment and collection of payments, scrutiny of claims, and data analysis, and other tasks consume a significant amount of time and effort.

When a car owner files a claim for vehicle insurance these days, he or she must first contact the insurance company to tell them of the problem, then proceed to the nearest police station to register a FIR and submit the insurance company the relevant application materials. The car owner will not be compensated until the degree of the damage to the vehicle has been certified by the company. According to studies, present automobile insurance claims procedures are often difficult and take an inordinate amount of time to compensate [2, 3].

Insurance reimbursement is time-consuming due to its complexity, and policyholders commonly have unresolved difficulties. Second, insurance companies spend a large amount of time each year, among other things, on the premium payments process, the collecting the records of claim, service providers of insurance, and audits of government. To ensure that all parties fulfil and comply with the contract's agreed-upon requirements, payment of claim and validation is a time-consuming including with huge amount of administrative costs and manually handling all the processes [4].

Blockchain has developed in prominence in recent years is used s technique through decentralization and trustlessness, for jointly keeping a reliable database, and it is now being used in a range of fields. Decentralization, trustlessness, non-tampering, clarity, and tracking are features of blockchain technology that are well suited for improving the structural adjustment in government, transparency in governance and service, stimulating the growth of intellect and trustworthiness. All over world same time government proposed various "Internet +" technologies with the help of blockchain in variety of government activities. With the help of cryptography and consensus mechanisms blockchain acquires decentralization for data based on peer-to-peer network this kind of decentralized storage system is termed as blockchain. Benefits of BC (Blockchain) technology include confidentiality, data integrity, and tracking of stored data. Insurance, as a risk-based industry, relies on data to stay afloat. A large amount of complete and correct information is required throughout the design of policy, assessment, and claims payout. With the insurance and associated industries, BC has the potential to create information conduits, industry integration can be improved, and increase insurance companies' data access capabilities. With the maturation of blockchain technology, a "blockchain + insurance ecological chain" can be built.

Blockchain has attracted much interest from academics [14], industry, and researchers in recent years, and it's been named one of the top five technology innovations of 2018 [15, 16]. According to [17], the daily output value of Bitcoin is 4.144 million as of September 17, 2020, with an anticipated transaction value of 158.932 million on the blockchain. Blockchain is categorized into three versions. First version of BC got published in 2009 and it was known as blockchain 1.0 which exclusively focused on digital money while nevertheless serving potentially malevolent worldwide participants [3, 18] was dependent on specific protocols. The second version came to existence in 2014 termed as blockchain 2.0 mainly focused on new methods to utilize smart contracts in different cases and domains with Ethereum [19], which offers client digital assets and partially complete capabilities [10], leading the way. Hyperledger projects published Blockchain 3.0 in 2017, with highly adaptable features (Fabric, Composer, and others) as well as user friendly it's a permissioned decentralized application system. Significant systems in logistics, certifications, and finance were established in the second generation of blockchain [20,21,22,23,49,50]. To construct the normality of blockchain technology, all three phases are complementary and assist one another [51]. According to [52,53], insurance is one of the most important forms of help available to communities in the event of an emergency, neutralizing expenses and assisting them. The sector's largest issue is detecting and protecting against counterfeit documents, as well as stopping the goals of phoney participants. Significant insurance companies have seen the impact of Blockchain technology, according to [54], with the majority of them investing in pilot systems. The B3i (Blockchain Insurance Industry Initiative), is one notable example started in 2016 to find out the benefits of blockchain to increase the success rate of data exchange in insurance organizations. As a result of providing a financial source for clients who are in a disastrous position, insurance companies are burdened with inefficiencies and piles of paperwork [56]. Centralized architecture was used by most insurance company in past for system development as shown in Figure 1(a) [53], Figure 1(b) shows the decentralized blockchain by using this architecture insurance company has a unique ability to better its entire value chain, which has always relied on the highest level of good confidence and belief, and define new insurance products for their customers [57] [58].

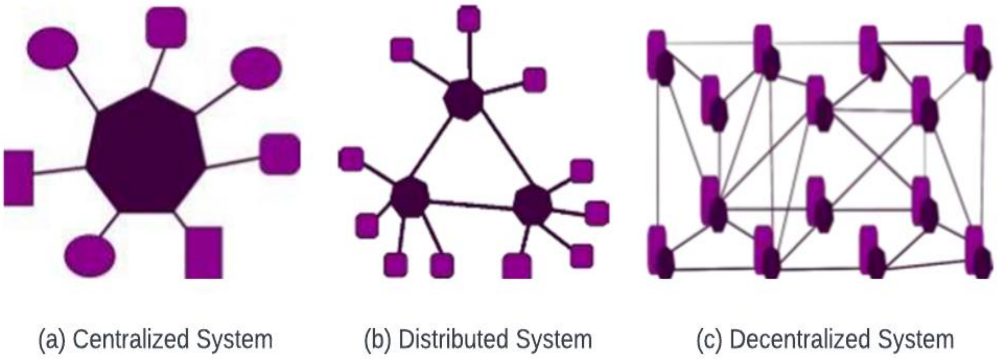


Figure 1: Existing system architecture.

We employed blockchain and smart contract technologies in this research to aid the development of Internet insurance in the following ways:

1.1.1 Mitigate difficulty of insuring effectively

All the insurance data, records of credit and information of individual vehicle will be collected one by one on blockchain which cannot be manipulated or fabricated once the financial sector adopts the BC technology. Adopting BC technology internet insurance can assess health and insurance information of clients online, effectively reducing the time and money for both clients and insurance companies. Reliability of internet insurance businesses' underwriting has considerably enhanced. Simultaneously, mode transfer from offline to online underwriting has been completed for internet insurance. Internet insurance has evolved into a business model rather than a sales channel.

1.1.2 Conducive to improving supervision

Internet crimes are growing increasingly common as a result of a lack of network oversight. On a technical level, blockchain technology has the ability to solve this problem in the future, notably in the operation and control of Internet insurance. Blockchain technology has two features: time tracking and changes that aren't editable. It also essentially verifies the data on the chain's authenticity. Network oversight can significantly reduce with the adoption BC technology in online insurance. On the one hand, the legitimacy of the operation of Internet insurance enterprises may be secured from the perspective of governmental control, thanks to the blockchain's timestamp features and the entire history of information and transaction data from the current block to the genesis block. It has ability to improve insurance industry's existing poor supervision and raise supervisory efficiency.

1.1.3 Suitable for resolving risk management issues

The most important goal is to avoid systemic risk. Moral hazard can be efficiently reduced using blockchain technology. The insured individual's credit history, health condition, asset registration, and use status will be established on the blockchain throughout the online insurance process, from underwriting to claim settlement, because the blockchain establishes a whole trust - free value network. Systemic risk can be avoided by allowing the insurer query all insured’s record on blockchain and assess the insured’s risk as the information is timestamped.

The second goal is to keep technical hazards under control. Existing computer technology used by internet insurance companies is insufficiently developed, resulting in crash occurs and phishing attacks. Security of internet insurance system is effectively safeguarded by BC technology due to its decentralized structure of ledger and nodes are also decentralized. Through communal maintenance likelihood of technical risks are reduced. The final goal is to keep information security threats under control. Customer information, firm internal information, and other data held by insurers, particularly Internet insurance providers, are extremely valuable. The BC technology assures secrecy of data held by Internet insurance businesses, this system is resistant to attempts to alter or destroy it by unscrupulous actors.

It ensures that Internet insurance companies run smoothly and, to some extent, lowers the cost of system maintenance.

1.1.4 Facilitating effective anti-money laundering

BC technology is used in internet insurance, which efficiently prevents money laundering. Every fund transaction has a unique time stamp, and records deposited into the blockchain must be recognized by the majority of blocks on the blockchain, ensuring the data's dependability and confidentiality to some degree.Simultaneously, transaction data cannot be changed whenever and wherever, and records and information on the blockchain cannot be destroyed whenever and wherever; as a result, time of transaction and the parties involved in transaction of each fund can be traced all the way back, in the context of criminal activities involving online insurance for financial fraud it is found to be really efficient method. It also serves as a more dependable platform for capital chain oversight.

In field of insurance applications, blockchain technology offers the following benefits:

- To some extent, claims settlement can be made more efficient by using blockchain smart contracts.
- To certain extent, the blockchain's security can help policyholders with their privacy and security concerns.
- Blockchain can address the issue of trust between policyholders and insurance companies.
- Blockchain can help to solve the problems related to records tampering in sales transactions.

Due to aforementioned characteristics of blockchain, it can aid both the insurers and the policyholder in establishing a long-term relationship. Insurers can use the blockchain mechanism to improve the efficiency of their operations and management, as well as precisely analyze operational risks. Selecting an insurance company that uses a BC claims settlement system can make claim settlement easier for the insured, allowing payment to be collected more fastly and rising the insured's faith in the company. Finally, this study may aid the insurance companies as a whole in improving its sustainability.

And the insured is involved in an accident and files a claim using the smart contract [25], the smart contract's automatic performance function is activated, and the claim payment is launched, when the system's predetermined requirements are met. To achieve auto settlement of claims, speed up claim settlement efficiency and save costs claim payment will be directly done to beneficiary's account.

The blockchain consensus mechanism presents a feasible solution to the issue of inefficient claim settlement. Smart contracts work on the idea of embedding terms of the contract in software so that the agreement cannot be broken and the cost of doing so is extremely expensive. Smart contracts provide predictability, uniformity, closure, verification, effectiveness, timeliness, and leastprice.

Whenever the information recorded in the blockchain meets the claim contexts, online insurance will initiate the clearing and settlement program immediately. The insurance payout agreement will be immediately terminated once the claim money is electronically given credit to the insured's chosen account. Using blockchain technology to automate claim settlement simplifies the payment process, eliminates several manual review processes, speeds up agreement completion, and lowers the price of agreement validation and completion.

1.2 Research goals

To accomplish the following research objectives, depending on blockchain and smart contract technology, an insurance claims system has been developed [59–89]:

- To design a framework for vehicle insurance claim process.

  Traceable: Allow for greater transparency in the process by allowing the car owner to quickly and easily file an insurance claim, resulting in a faster settlement. BC technology make’s sure that all have the equal right to know and select, as all the participants has equal access to data. Regards to the decentralized storage verification mechanism as modifications to the records will be in sync on the chain. In future if a disagreement arises policyholders’ rights are protected due the presence of evidence to prove.
- Eliminate multiple claims for one accident.
- Privacy: Using decentralized data access, you can achieve data privacy and security. Information that clients selects to share only will be stored behind the partnership chain. Because of signature secret keys, cryptographic algorithms, and protect multi-party technology, blockchain technology would allow authorized consumers only to connect record, as well asensures that the blockchain alliance member database's basic data and privacy are not leaked. To preserve users' privacy, the insurance contract's content is password-protected. Party can view only personal contract and not the key it is in the party's possession. Smart contracts will be used to effectively integrate the insurance contract, with agreement overview, request, amendment, as well as other details occurring and being stored in the block.
- Non-repudiation: To avoid fraud, provide time stamped transactions. The data will be permanently preserved once it has been verified and uploaded to the blockchain, and the blockchain's inherent time stamp mechanism may note time of transaction. It will very difficult to change the information as it is required to control more than 51% of system nodes.

1.3 Existing vs proposed system

This section highlights the novelty of the proposed system compared to existing system in the insurance domain. Novelty of proposed system is presented by comparing it with existing system as follows

Table 1: Existing system vs proposed system

Existing system	Proposed system
Actual value is transferred between parties by the centralized administration.	Through a cryptographic technique that offers a shared information source on remote nodes, trust is generated.
Contract development, fund transfer, and complex manual information/data review on a physical document.	Real-time access to and analysis of all financial papers enables the immediate start of shipments.
Delays brought on by manual communication and changes made to the contract conditions by all parties	Transaction time is shortened through contract execution on decentralized nodes and real-time status updates.
Due to centralized restrictions, there is no transparency about ownership and location proof.	Transparency brought on by decentralized management.
The transaction charge is increased by the manual process.	Transaction fees were less expensive thanks to automated settlement.
Multiple copies of documentation make it difficult to control modifications that are made.	Multiple parties participating in the transaction process can all examine the document simultaneously in real time without encountering any problems.

The manual procedure produces various platforms for each side, which greatly increases the likelihood of miscommunication and fraud.	With a single platform, there is no room for fraud as everyone receives the same communications.

2 Preliminary

2.1 Blockchain technology

BC technology is a distributed ledger that tracks the history of transactions and is publicly verifiable, distributed, and unchanged. A blockchain, as the name implies, is a collection of blocks that each carry details of transaction and are linked together by a hash of the blocks before and after them to form a chain. The blockchain system is made up of nodes, each of which keeps a replica of the chain's data and connects to other nodes via point-to-point connection. A header, an ID for the past and subsequent blocks, a date, and a set of transactions are all included in each block. Blockchain (BC) is a decentralized technology that allows for the creation of brand-new technological operations and business plans [89]. Blockchain integrates earlier developed technology such as electronic certificates, cryptographic hashes, and decentralized consensus techniques [90]. The underlying digital foundation that underpins apps like bitcoin is known as blockchain. In a cooperative network, the technology improves the process how the transactions are stored and assets are tracked. Practically any other sort of asset that can be exchanged between peers and stored safely and confidentially with no need for third-party authentication can be used as assets. This is because cryptography, network consensus mechanisms, smart coding, and teamwork enforce confidence and provide confirmation without the need for controlling intermediates such as governments and banks [91]. The scientific research of data structure is the foundation for current algorithms, in which nodes are used for datagrams and data warehouses, and they interact with each other via agreed-upon node protocols [92].Blockchain can be discussed alongside other similar techniques because they also can be outlined and reviewed based on empirical data structure studies, with the goal of serving as the core component of existing techniques in which endpoints have been used for datagrams and information repositories, interacting with one another using approved methods. A transaction was indeed demanded in Blockchain, and it is then conveyed to the point-to-point network. When a transaction is authenticated and verified, it is linked to the blockchain's existing blocks, completing it. Figure 2 illustrates this point.

2.2 Features of blockchain

Immutability and security: Blockchain provide a safe and consistent method of storing and retrieving information among nodes in blockchain systems due to its immutability [93]. Unchanging transactions are protected from malicious users' illegal access. Participants can add new transactions but not delete or amend existing ones, making it easier for all nodes to keep records of previous transactions [94]. Information can never be modified after it has been written and kept in the ledger [95]. If a transaction has a mistake, a new transaction must be made, and both transactions arepresent. Because all nodes have complete details for identification, confirmation, and validity, as a result, it reduces reliance on a central integrity entity and the risk of a central entity malfunctioning or manipulating data. [94].

Transparency: The blockchain network imposes confirmation and acceptance testing through a consensus approach, in which any participant can appear and openly begin and append transactions after complying with the blockchain's rules.Since all transactions must be approved by their intended producers, and once a block has been approved, miner nodes send this same block to all of the other nodes within the network, blockchain's consensus, confirmation, and admittance methods ensure network members' confidence [95]. All transactions are made public to all nodes who are part of the networkin this situation, as they are in a public blockchain, but all data in a private blockchain is only accessible to authorized nodes.

Verifiability: Outsiders and insiders can verify transactions conducted and maintained via blockchain technology thanks to cryptography and consensus mechanisms. To be genuine, at least 50 % of participants plus one has to agree on the transaction's validity.

Authenticity: The use of consensus mechanism in blockchain applications ensures the legality of transactions. Furthermore, each block in the blockchain contains past and successive hashed IDs, as well as the producer's and responder's digital signatures.

Ownership and accountability: ownership and accountability can be provided based on the connection between blocks, integrity of transactions, and the approval of the original creators in blockchain-based systems. In addition, participants are aware of a block's or transaction's provenance.

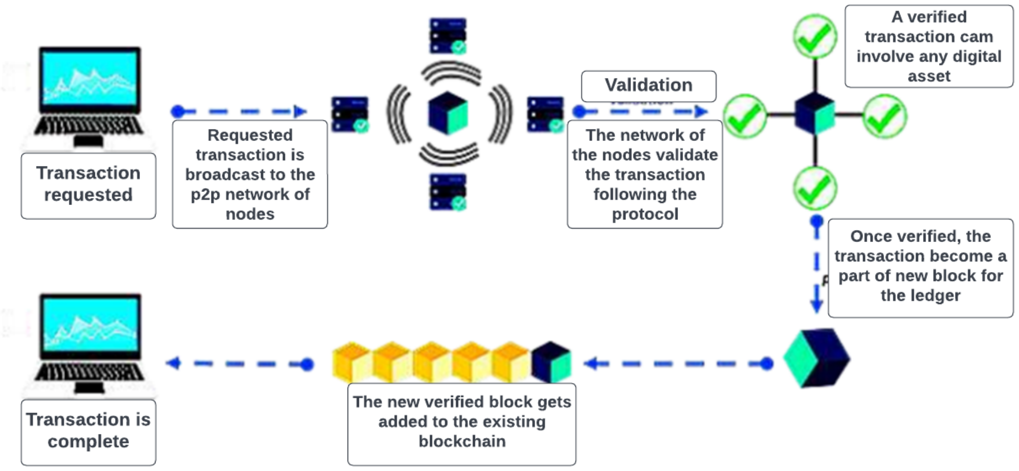

Transaction requested
Requested transaction is broadcast to the p2p network of nodes
Validation
The network of the nodes validate the transaction following the protocol
A verified transaction cam involve any digital asset
Once verified, the transaction become a part of new block for the ledger
The new verified block gets added to the existing blockchain
Transaction is complete

Figure 2: A visual representation of how blockchain works, source: Edureka.

2.3 Blockchain components

Assets:by definition, it's something precious to a company, it enables the transfer of well almost anything of commercial benefit across a blockchain platform. An asset can be intangible: money, stocks, intellectual property rights, certificates, and personal data, or tangible: foodservice at a hotel, real estate investments, depending on the blockchain system. Compared to conventional -assets such as Apple Stock, which have a paper-based right of ownership, a blockchain asset is entirely digital and completely owned by the participant, requiring no third-party agent or agency for transfer or sale.

Transactions: A blockchain transaction is a sequence of time-stamped events that lead to the production of blocks in the ledger. Participants can stay anonymous since transactions are saved using private or public keys; however, third parties can access and verify identities. Most transactions and perhaps other data are publicly reviewed before being put into the ledger using Ralph Markel's tree, as depicted in Figure 3 [96], to ensure system trust and harmonization. The Merkle tree [96] is a huge binary tree data structure which improves in data consistency validation and guarantees, allowing for speedier security authentication in big data applications. The value from each child node is hashed by the parent node.

Algorithm for achieving consensus: A consensus method aids decision-making in decentralized or distributed systems [97]. In a blockchain network, the consensus algorithm is an administrative system in which the majority of untrusted parties agree on the rules to be followed and the best alternative for everyone.Quorum structure, truthfulness, decentralized governance, authentication, nonrepudiation, performance, and byzantine fault tolerance are all characteristics of the blockchain consensus algorithm [97].Whether or not a block is added to the blockchain is determined by this decision.The consensus algorithm plays an important function through enabling collaboration and cooperation, as well as assuring that all members have equal rights and recognition and encouraging constructive participation.For a thorough discussion of consensus algorithms, see [98].

Functions in cryptography: Cryptographic Functions employ complicated mathematical computations to transform data into information that is completely useless inside the hands of the wrong people. This blockchain feature enables potentially harmful blockchain network participants to build and append blocks to the chain, as well as conduct secure network operations. Every block of the blockchain contains the immediately preceding block's hash, as well as transaction records and a time and date. Unsymmetrical (public-key) cryptography, in contrast to symmetric key encryption, encrypts with a publicly shared public key and thereafter decrypts with a private key. Blockchain procedures are also secured via hashing. Prior to a transaction, the immediately preceding block's data is hashed and saved. The public key of the transaction creator is hashed and used to establish transaction information about the identity [99, 100, 98].

Distributed ledger: In the marketing world, a ledger is absolutely vital because it stores all records for online and offline activities, as well as clients and their credentials. A ledger is centralized in a typical business IT environment.

The blockchain ledger, on the other hand, is decentralized at its essence, and regardless of the nature of blockchain, it can be accessed by a small number of authorized parties (private) or by all participants (public) (public). Auditability, security, transparency, and accountability are all enforced by the distributed ledger, along with other BC characteristics.

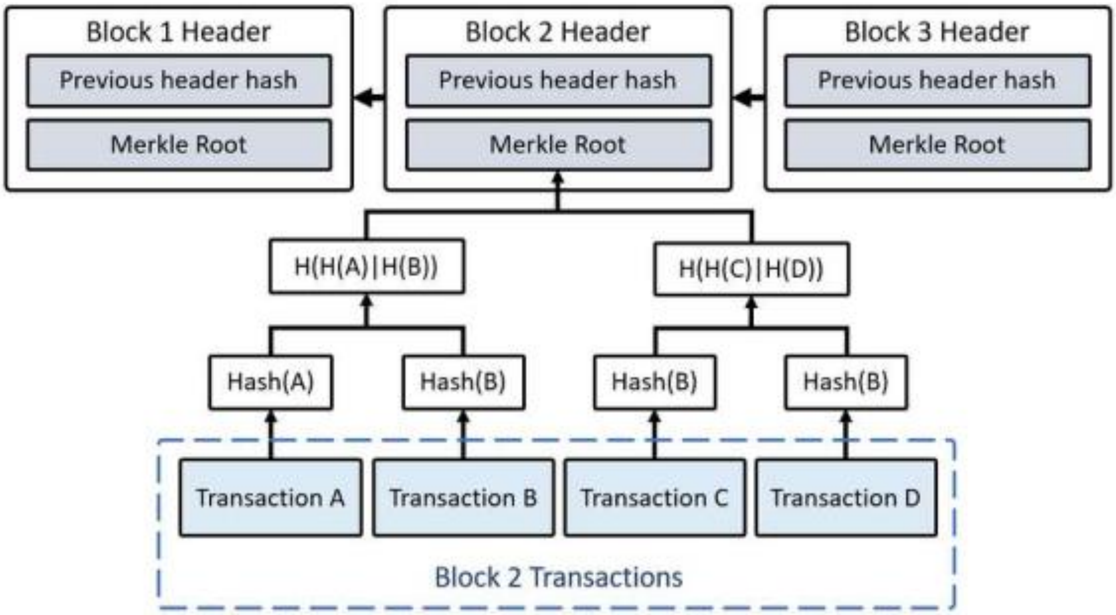


Figure 3: Merkel tree is used to illustrate blocks in blockchain.

2.4 Blockchain types

Public blockchain: This is accomplished without the need for permission on a public blockchain by enabling everybody to participate in the consensus process. Everyone who is a member of a public blockchain can interpret, try writing, and exchange on the network [101,102]. There is no single trustworthy organization in responsible of network supervision and control, hence it is decentralized. This kind of blockchain is safeguarded by encryption, which incentivizes miners to verify it. Miners, who can be anyone on the blockchain, consolidate and propagate transactions [101, 102]. Because no member of the blockchain is trustful, the public blockchain relies on computer systems and brute strength tactics to validate transactions. As a consequence, the miner who achieves all of the right answers at the conclusion of the process rewarded. The most extensively utilized public blockchains are Bitcoin and Ethereum [101, 102].

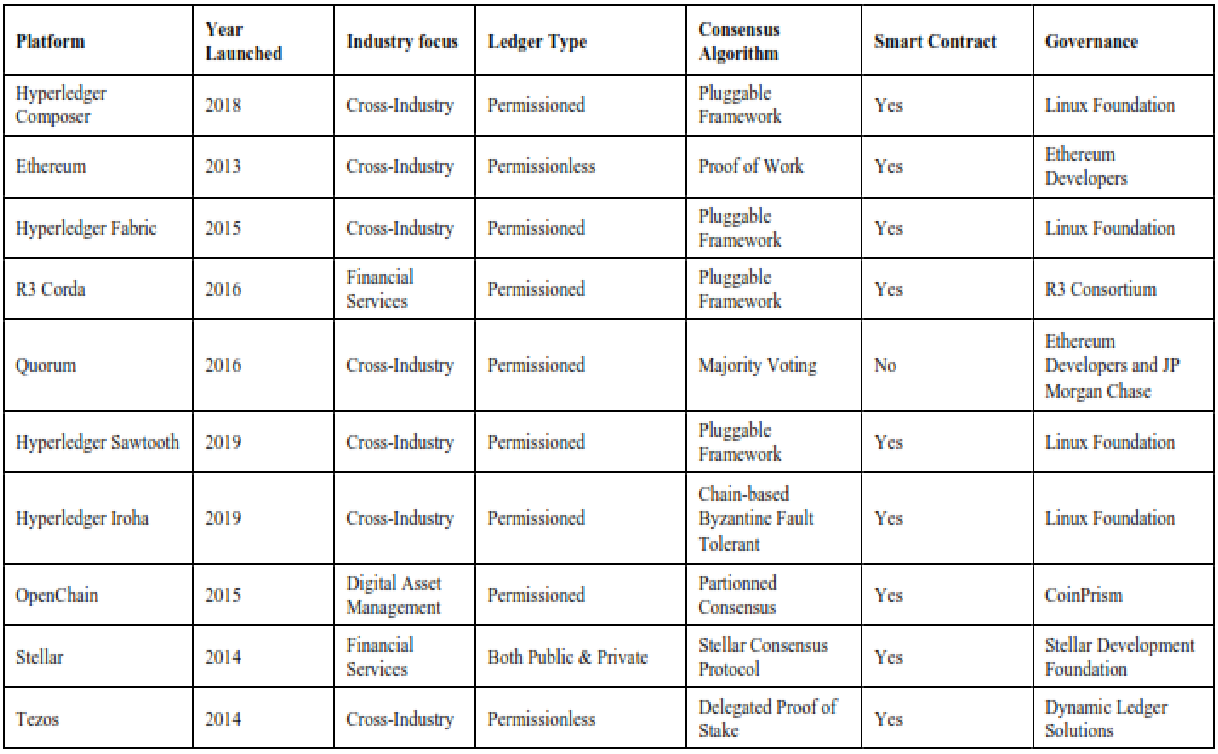
Private blockchain: this is a permissioned blockchain in which network users are restricted from accessing certain areas of the blockchain. This imposes a centralized control mechanism and allows just a few network participants to modify the network. Because it meets with standards and regulations such as KYC and AML, this sort of blockchain is commonly employed in the banking industry. [102, 101].

Consortium blockchain: this is a quasi-decentralized, permissioned blockchain. Unlike a public blockchain, here it needs permission to enter and only allows a small number of clusters to manage and administer the network. The public clients may indeed be given restricted access to the blockchain through API, with only the most basic of queries possible. This type of blockchain maintains the inherent safety for data of public blockchain while also providing increased network control. Platforms like R3, Quorum, and Corda [102, 100] are examples.

2.5 Blockchain platforms

The sections that follow go over some of the most popular blockchain platforms. Table 2 summarizes this information.

Table 2: The most popular blockchain platforms, source: www.ijacsa.thesai.org 449



Platform	Year Launched	Industry focus	Ledger Type	Consensus Algorithm	Smart Contract	Governance
Hyperledger Composer	2018	Cross-Industry	Permissioned	Pluggable Framework	Yes	Linux Foundation
Ethereum	2013	Cross-Industry	Permissionless	Proof of Work	Yes	Ethereum Developers
Hyperledger Fabric	2015	Cross-Industry	Permissioned	Pluggable Framework	Yes	Linux Foundation
R3 Corda	2016	Financial Services	Permissioned	Pluggable Framework	Yes	R3 Consortium
Quorum	2016	Cross-Industry	Permissioned	Majority Voting	No	Ethereum Developers and JP Morgan Chase
Hyperledger Sawtooth	2019	Cross-Industry	Permissioned	Pluggable Framework	Yes	Linux Foundation
Hyperledger Iroha	2019	Cross-Industry	Permissioned	Chain-based Byzantine Fault Tolerant	Yes	Linux Foundation
OpenChain	2015	Digital Asset Management	Permissioned	Partionned Consensus	Yes	CoinPrism
Stellar	2014	Financial Services	Both Public & Private	Stellar Consensus Protocol	Yes	Stellar Development Foundation
Tezos	2014	Cross-Industry	Permissionless	Delegated Proof of Stake	Yes	Dynamic Ledger Solutions

Ethereum: The blockchain community and developers can use this platform to create and deploy smart contracts applications. The Ethereum Distributed Environment is open-source and allows for the deployment of tokens, cryptocurrency, social apps, wallets, and more. Blockchain technology can be used in a diverse range of companies, not simply banking, thanks to Ethereum's architecture. This platform is made up of a number of different components as follows

- Smart contracts, defined in the Solidity programming language, are used to control all occurrences in Ethereum.
- On the Ethereum network, Ether is the backbone of transactions and cryptocurrency.
- In this platform clients are the people who create and mine the Ethereum blockchain. Geth, Eth, and Pyethapp are some examples.
- The EVM is a blockchain engine which ensures smart contracts for work. EVM's programming language is bytecode, which necessitated the creation of a variety of different smart contract authoring languages, such as Solidity.
- Etherscripter is an user interface that allows you to create Ethereum smart contracts. The drag-and-drop technique enables for the automatic creation of backend codes in LLL, Serpent, and XML in only a few simple steps.

Hyperledger: this is a global partnership led by The Linux Foundation and comprised of industry professionals from financial, accounting, production lines, IoT, technologies, and industries. This platform is an open-source project led by a community dedicated to putting together a set of solid foundations, libraries, and tools for building and implementing organization blockchain systems. Permissioned (private) frameworks exist alongside permissionless (public) frameworks [104].

Corda: This blockchain platform was released by R3 in 2016 as an open-source blockchain technology with widespread support from developers and organizations. This platform has a characteristic that no other blockchain has. It's a restricted blockchain network where only known members can share information. The purpose of Corda was to promote trust, openness, protection, and privacy.

Quorum: This platform of blockchain is a business-oriented blockchain. It's an improved version of the open-source Ethereum client 'geth' that caters to business demands. It is an open-source project that addresses the functionality, security, and access control concerns of businesses. Quorum satisfies the requirements of corporate applications, which include privacy, performance, and permissioning, as well as transaction secrecy, scalability, and speed, as well as authorization.

3 Literature survey

Blockchain technology has risen to prominence as a cutting-edge disruptive technology. As shown in Table 2, several efforts have been made to broaden the scope of blockchain's usefulness. However, there are only a few works in the insurance industry. As a result, we look at the various opportunities and threats that this endeavor presents. Claims and fraud are the most crucial business processes in the insurance industry that may be improved or re-engineered, according to industry and academic studies. Blockchain's nature, as well as aspects like as cryptography, consensus algorithms, decentralization, and others, make it a perfect remedy for the finance industry. The challenge of data recognition and intelligent data transfer is handled by hashing the identities of members in the blockchain network. The blockchain P2P model has the potential to establish a new range of insurance products while eliminating the need for trusted middlemen. This analysis was carried out in order to find the possibilities that would best position us to begin our future projects. Security and data privacy, scalability, legislation, and taxation are among problems that blockchain technology now faces. Despite the fact that the separate technologies on which blockchain is built are mature, their integration introduces vulnerabilities. Blockchain will become a very strong tool for addressing many of the technical issues that the insurance business is currently facing.

Table 3: Literature survey summary

Authors, Title of Paper, year of publication	**Gaps identified**	**Tools Used**	**Methodology used**	**Major Results**
Anokye Acheampong AMPONSAH *et al. [58] ,2021	• Blockchain 3.0 has yet to be put to good use in the insurance industry. • As previously said, all insurance sector initiatives remain in their infancy. • Because not all data is required by all nodes in private blockchain systems, decentralization of the entire ledger could be theoretically insignificant, but it could worsen storing, scaling, and technical problems.	Hyperledger fabric	Locating Studies or Data Extraction, Data Screening and Selection, Descriptive Analysis	• The insurance industry, as large as it is now, has realised that blockchain technology may be used to improve essential internal procedures such as submission and processing of claim, detection and prevention of fraud, and so on.
Mayank Raikwar∗ et al. [59], 2018	• Transaction management time, clearing and settlement time, and security are all major concerns in the insurance process.	Hyperledger fabric, solo consensus algorithm	Implemented an insurance company's operations into smart contracts and stored the outcomes in a blockchain-enabled distributed platform.	The Confirmation time is related to the network size.
Aarti Patkiet al. [63], May 2020	• There is only one point of failure. • Deploying suitable legal framework is a huge task. • Time-consuming and costly KYC (Know Your Customer) process can be completed more quickly and at a lower cost on BCT.	--	Top Indian banks, FinTech organizations, and banking consultants were contacted for primary research. We contacted each of the 25 execs. The information was gathered through organized questionnaires and interviews with senior executives. Secondary data was gathered from Deloitte news bulletins and	Addresses both primary and secondary research aimed at learning more about BCT and how it's used in banking.

			publications, as well as research articles published in research journals.	
D. Popovic et al. [60], 2020	• There are no well-established standards or platforms. • The inability to initiate a claim on more complex insured occurrences due to a lack of trustworthy third-party data.	enterprise risk management (ERM)	**Project mobilisation –** Form a team to carry out the solution you've proposed. Plan more thoroughly. **Delivering capabilities:** Placing processes and mechanisms in position to help the solution to be reality. **Launch -** After testing, deploy the application from a development platform to a real system.	Studying, analysing, and utilising blockchain as a practical reference for insurance sector practitioners.
Vukolic and Marko [61], "Rethinking permissioned blockchains", 2017	• Smart contracts operate progressively, all nodes execute consensus procedures are tricky in all smart contracts, the framework is rigid, and smart-contract execution is non-deterministic, causes major challenges on current blockchain platforms, particularly recent permissioned systems.	Hyperledger fabric	• Design constraints Of Permissioned Blockchains • Using Hyperledger Fabric to Overcome Limitations	A study at the constraints that various permissioned blockchains have.
C. D. Clack, V. et al. [62], 2016	• The duties and responsibilities of those who are able to function under a contract (e.g. designated signatories) • The ability to indicate that if a specific phrase is incorporated or modified, the agreement must be forwarded to a third party for special approval.	Grigg's Ricardian Contract triple	• Presenting the essential requirements for smart legal agreements, • A smart contract's and a smart legal contract's abstract fundamental structure. • The development environment for a structured format for smart legal contracts storage and communication.	• Identification of essential requirements • Description of number of key design options.

Mitt, Sven, [71],"Blockcha in Application - Case Study on Hyperledger Fabric", 2018	• Lack of distributed cross-chain transactions • Support and documentation and inability to support today's fast delivery pace.	open-source Hyperledger Fabric	A network of nodes running Hyperledger Fabric is created and parking spot application is deployed into the network as a smart con- tract.	• On transactions, Hyperledger Fabric supports ACID features (atomicity, consistency, isolation, and durability). • Using validation and business rules in smart contract that provides strong consistency enables strong trust toward the correctness of data and the entire system.
Guy Zyskindet al.[72], 2015	Third-parties collect and control massive amounts of personal data.	mobile software development kit (SDK)	The system is made up of three components. • Users of mobile phones who want to download and use applications; services • Developers of such services who need to process personal data for operational or economic purposes, as well as • In exchange for incentives, nodes are entities methods of managing the blockchain and a distributed private key-value data store.	• Resolved users' privacy concerns during using third-party services. • The emphasis is on mobile platforms. • Applications capture high-resolution personal information on a continuous basis without the user's knowledge or consent.

4 Proposed model

Our approach is based on the concept of implementing insurance provider procedures as consensus mechanism and storing the results in a distributed blockchain platform [59].

4.1 The model's entities

The Agent, who works on behalf of the client and handles the customer's queries to the blockchain network, and the Customer, whom was protected by insurance and requires insurance policies, makes claim demands, and receives reimbursements, are the two main entities in our concept. An agent can work with many clients.

4.2 The model's components

The core elements of our architecture are a decentralized blockchain ledger B which keeps records of all transactions' execution results in (Key, Value) format, a database DB (preferably encrypted) which keeps track of all clients' insurance contracts and transaction results in (Key, Value) feature, list of endorsers ESC who authenticate the transaction situations of blockchain network, and a set of orderers O who order the transactions sequentially and develop transaction history. Individuals are authenticated and access is controlled using cryptographic procedures.

4.3 Framework for insurance

The insurance structure is made up of assets that allow the network to interchange almost anything with monetary worth. The framework's rules for transactions are governed by smart contracts. In the insurance blockchain network, a block is formed when peer nodes in the validator set V reach consensus on a group of transaction results [105]. Every smart contract has endorsement (or verification) logic that defines the conditions under which it can execute a transaction. The endorsement logic is performed out by a group of endorsers ESC who examine whether contract criteria are met using the blockchain.

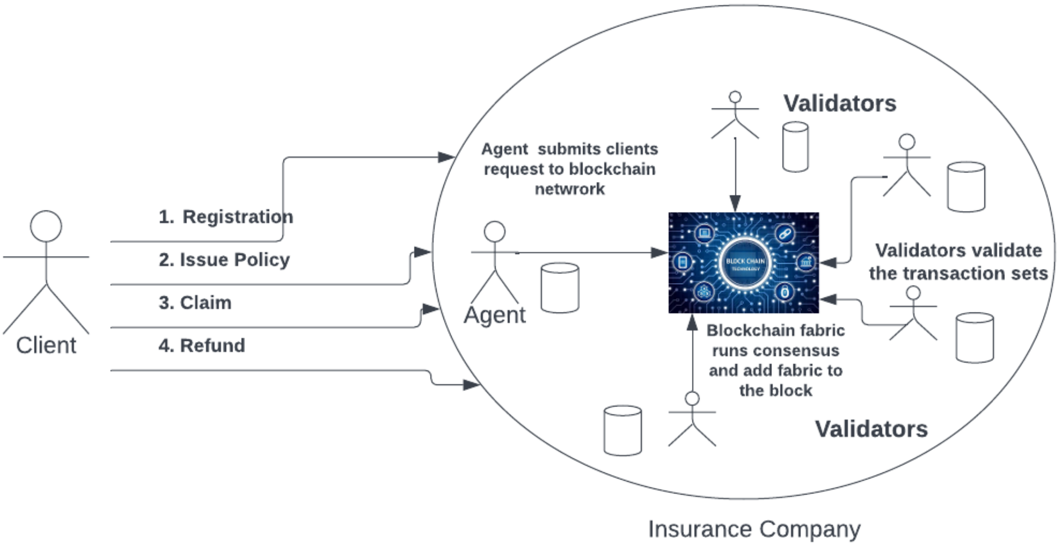

Validators
Agent submits clients request to blockchain netwrork
1. Registration
2. Issue Policy
3. Claim
4. Refund
Client
Agent
Validators validate the transaction sets
Blockchain fabric runs consensus and add fabric to the block
Validators
Insurance Company

Figure 4: Insurance blockchain framework system model.

4.4 Insurance business model

We'll look at a situation in which the major processes (transactions) are normal insurance operations such as registration of client, assignment of policy, payment of premium, submission of claim, processing refund, and so on. Each transaction is recorded on the blockchain, ensuring that clients do not unfairly accuse the insurer and that the insurer is held responsible for all of its activities. The framework's core workflow is depicted in Figure 5.

Each smart contract is unique. SC_j has a group of endorsers called ESC_j who sign off on the contract's transactions. The word object relates to the client's or policy's attributes. The format of an object is determined during the instantiation of a smart contract. Function f is used to build an object from its properties and function f is used to produce composite keys (primary) from the ID (s). The key's function is to obtain specific object(s) from the database that correspond to the ID(s). We also use partial composite keys (not primary) to get a set of objects from the database in our system. As follows, we go over each contract in depth.

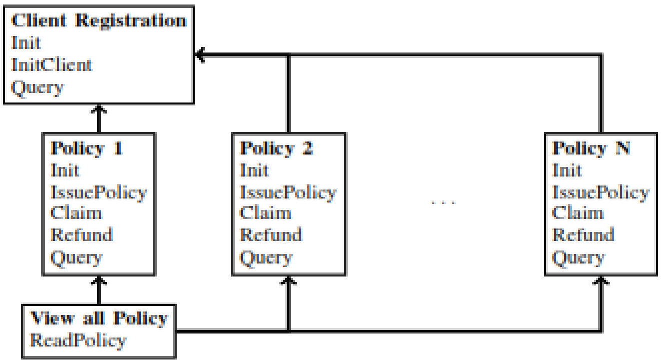
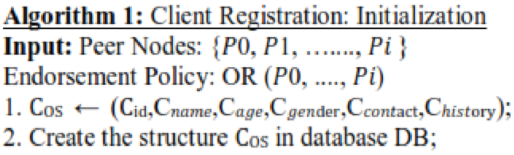


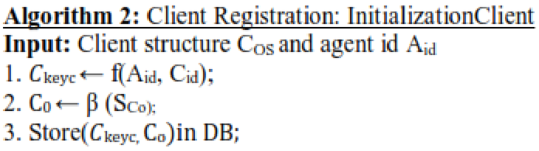
Figure 5: Smart contracts for insurance processes.

Client registration: Clients are registered in the insurance system via smart contracts. During the initialization of the database DB, a structure for the client object (C_o) is created (Algorithm 1). Client attributes such as $C_{id's}$ unique id are utilized as keys, while other client attributes will be used as values.



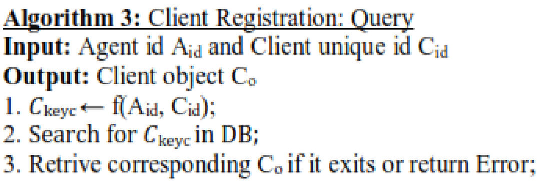
```
Algorithm 1: Client Registration: Initialization
Input: Peer Nodes: {$P0, P1, \ldots\ldots, Pi$ }
Endorsement Policy: OR ($P0, \ldots, Pi$)
$C_{OS} \leftarrow (C_{id}, C_{name}, C_{age}, C_{gender}, C_{contact}, C_{history})$;
Create the structure $C_{OS}$ in database DB;
```

An agent creates composite key C_{keyc}, and client object Co is constructed using C_{keyc}used to register a customer (Algorithm 2).



```
Algorithm 2: Client Registration: InitializationClient
Input: Client structure $C_{OS}$ and agent id $A_{id}$
$C_{keyc} \leftarrow f(A_{id}, C_{id})$;
$C_0 \leftarrow \beta\ (S_{Co})$;
Store($C_{keyc}, C_o$)in DB;
```

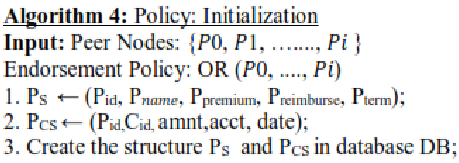
To obtain specific customer information, an agent A of insurance must produce a composite key C_{keyc} (Algorithm 3).



```
Algorithm 3: Client Registration: Query
Input: Agent id $A_{id}$ and Client unique id $C_{id}$
Output: Client object $C_o$
$C_{keyc} \leftarrow f(A_{id}, C_{id})$;
Search for $C_{keyc}$ in DB;
Retrive corresponding $C_o$ if it exits or return Error;
```

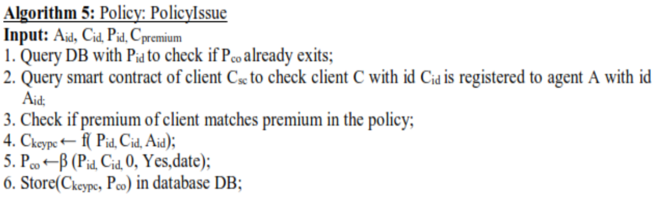
If an agent wishes to access all of his or her customers, he or she can create a partly composite key PCkey with just his or her personal id Aid and use it to scan the database DB.

Policy: Policy issuance, claims, and reimbursements are all part of a smart contract. Structures of policy and policy clients P_S, P_{CS} are created in database DB at startup (Algorithm 4), where amnt, acct, and date on which the amount claimed, indicator of claim acceptance (yes or no), and submission date of claim, respectively.



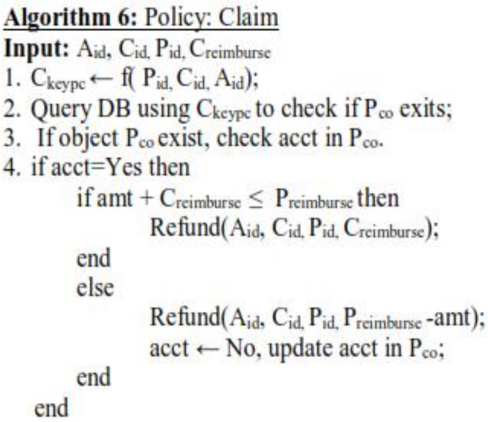
```
Algorithm 4: Policy: Initialization
Input: Peer Nodes: {$P0, P1, \ldots\ldots, Pi$ }
Endorsement Policy: OR ($P0, \ldots, Pi$)
$P_S \leftarrow (P_{id}, P_{name}, P_{premium}, P_{reimburse}, P_{term})$;
$P_{CS} \leftarrow (P_{id}, C_{id}$, amnt,acct, date);
Create the structure $P_S$ and $P_{CS}$ in database DB;
```

Client c selects policy P (id P_{id}) from the available policies and pays a premium $C_{premium}$ to the agent A id (A_{id}) in the policy issuing process (Algorithm 5). If the transaction passes all of the regular tests and verifications, the database creates and saves a policy client object called P_{co}.



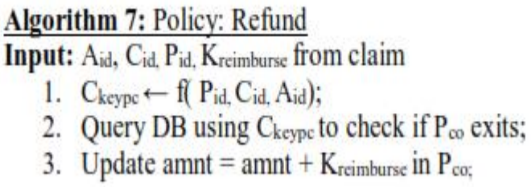
```
Algorithm 5: Policy: PolicyIssue
Input: $A_{id}, C_{id}, P_{id}, C_{premium}$
Query DB with $P_{id}$ to check if $P_{co}$ already exits;
Query smart contract of client $C_{sc}$ to check client C with id $C_{id}$ is registered to agent A with id
    $A_{id}$;
Check if premium of client matches premium in the policy;
$C_{keypc} \leftarrow f(P_{id}, C_{id}, A_{id})$;
$P_{co} \leftarrow \beta\ (P_{id}, C_{id}, 0$, Yes,date);
Store($C_{keypc}, P_{co}$) in database DB;
```

To handle a claim, client c transmits his credentials to the appropriate agent A (Algorithm 6).The refund process is started if all of the essential conditions are verified. If the claim is accepted, the acct parameter is set to true.



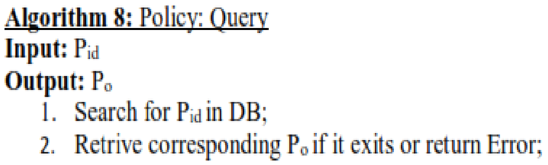
```
Algorithm 6: Policy: Claim
Input: $A_{id}, C_{id}, P_{id}, C_{reimburse}$
$C_{keypc} \leftarrow f(P_{id}, C_{id}, A_{id})$;
Query DB using $C_{keypc}$ to check if $P_{co}$ exits;
If object $P_{co}$ exist, check acct in $P_{co}$.
if acct=Yes then
        if amt + $C_{reimburse} \leq P_{reimburse}$ then
              Refund($A_{id}, C_{id}, P_{id}, C_{reimburse}$);
        end
        else
              Refund($A_{id}, C_{id}, P_{id}, P_{reimburse}$ -amt);
              acct $\leftarrow$ No, update acct in $P_{co}$;
        end
    end
```

The claim process starts off the refund process. During the reimbursement process, the total amount claimed amnt in the P_{CO} is changed in DB.



```
Algorithm 7: Policy: Refund
Input: $A_{id}$, $C_{id}$, $P_{id}$, $K_{reimburse}$ from claim
    1. $C_{keypc} \leftarrow f(P_{id}, C_{id}, A_{id})$;
    2. Query DB using $C_{keypc}$ to check if $P_{co}$ exits;
    3. Update amnt = amnt + $K_{reimburse}$ in $P_{co}$;
```

Agent A can acquire information about his or her clients who have purchased a specific insurance P_{id} by searching the DB with a key P_{Ckey} created by A_{id} and P_{id}. This basically retrieves $\{P_{coi}\}i \in \{0,...,N\}$through the database.An agent can indeed retrieve information about each particular policy issued by the insurance company by using the Search queries method in the policy smart contract.



```
Algorithm 8: Policy: Query
Input: $P_{id}$
Output: $P_o$
    1. Search for $P_{id}$ in DB;
    2. Retrive corresponding $P_o$ if it exits or return Error;
```

4.5 Notations

The notation of the proposed scheme as follows

C_o	client object
DB	database
C_{id}	client unique id
C_{Keyc}	Composite key c
I_A	Insurance Agent A
A_{id}	Agent id
PC_{key}	Partial composite key
P_S	Policy structure
P_{CS}	Policy client structure
amt_c	claimed amount
C_{AI}	Acceptance Indicator of Claim
C_{SD}	Submission Date of Claim
P_{id}	Policy id
C	client
P_{CO}	Policy client object
C_{os}	Client Structure
C_{SC}	Client smart contract
P_o	Policy Object
AML	Anti-Money Laundering
EVM	Ethereum Virtual Machine

4.6 Transactions in the framework

Client c sends a transaction request to agent A on the proposed blockchain network. The request includes the smart contract technique and client attributes required for the function to operate. Agent A signs the transaction, which is then approved by the smart contract's endorsers. After the transaction has been validated, Agent A presents it to the ordering nodes O in order to arrange the transactions chronologically. With all of the transactions they have received, the peer nodes run the core consensus function, attaching the new records to the blockchain.

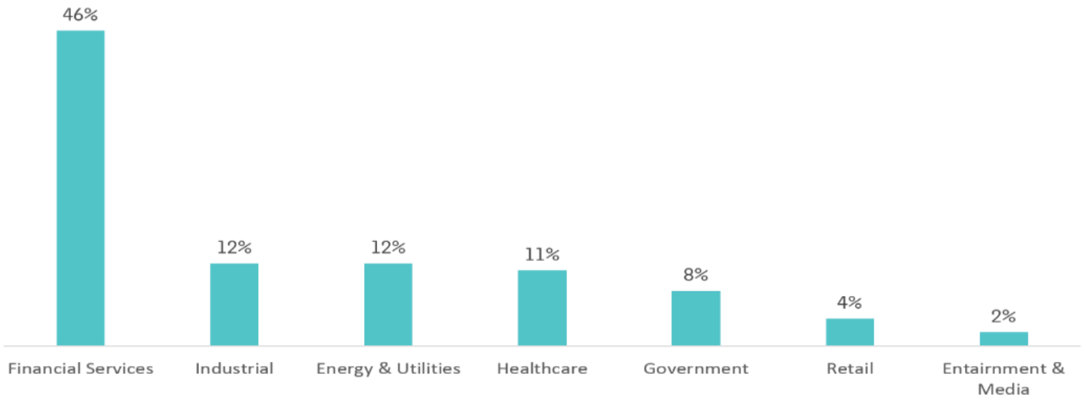
4.7 Potentail of blockchain in financial services

While many insurance businesses still rely on a conventional setup and have little knowledge of their clients, they are unable to provide the services that customers expect. The biggest drawback from the perspective of an insurance provider is the enormous gap between this method of service supply for clients and the real service provision. However, a lot of people are unaware of blockchain technology can improve the security and efficiency of the insurance process.

The process of filing insurance claims can be streamlined using blockchain technology. It is possible to rapidly and securely verify claims and process them. Additionally, blockchain is impervious to corruption and tampering because it is a distributed system.

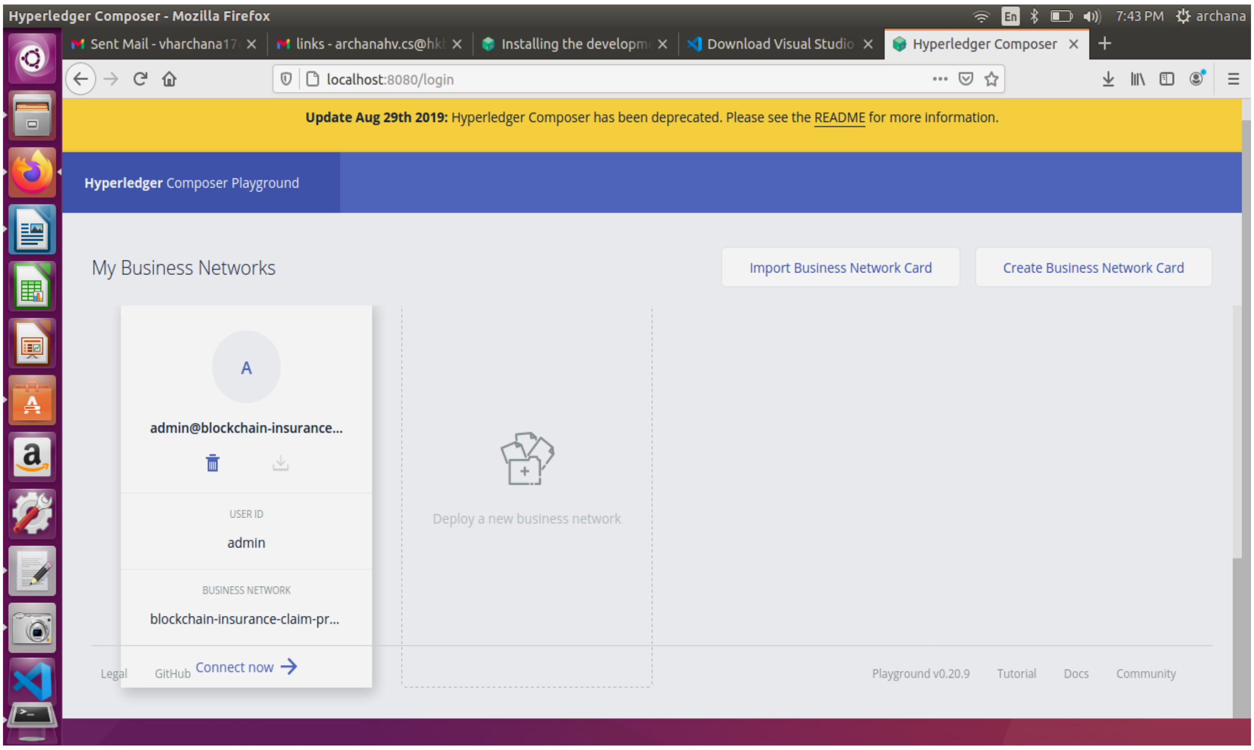
The financial industry has a larger range of applications for blockchain [106], including:

- Smart contracts are used for real-time trade settlement at decreased cost, the issue of commercial paper, and the settling of delivery and payment.
- Eliminating errors caused by manual auditing and shortening the trade finance process with minimal middlemen in international trade
- Online application and claim settlement for insurance.



Source: Credit Suisse

Figure 6: Potential of Blockchain among various industries, in %, 2018, source: Credit Suisse



Update Aug 29th 2019: Hyperledger Composer has been deprecated. Please see the README for more information.

My Business Networks

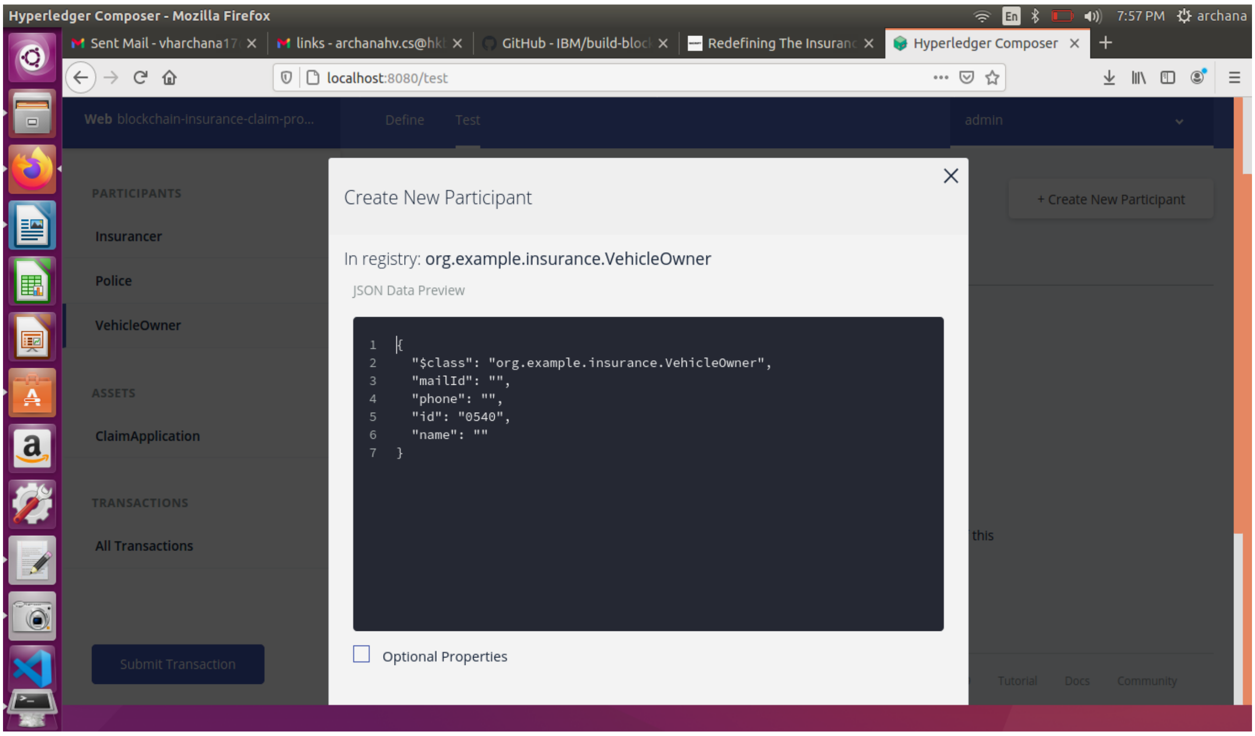
A

admin@blockchain-insurance...

USER ID
admin

BUSINESS NETWORK
blockchain-insurance-claim-pr...

Figure 7: My business network page.



Create New Participant

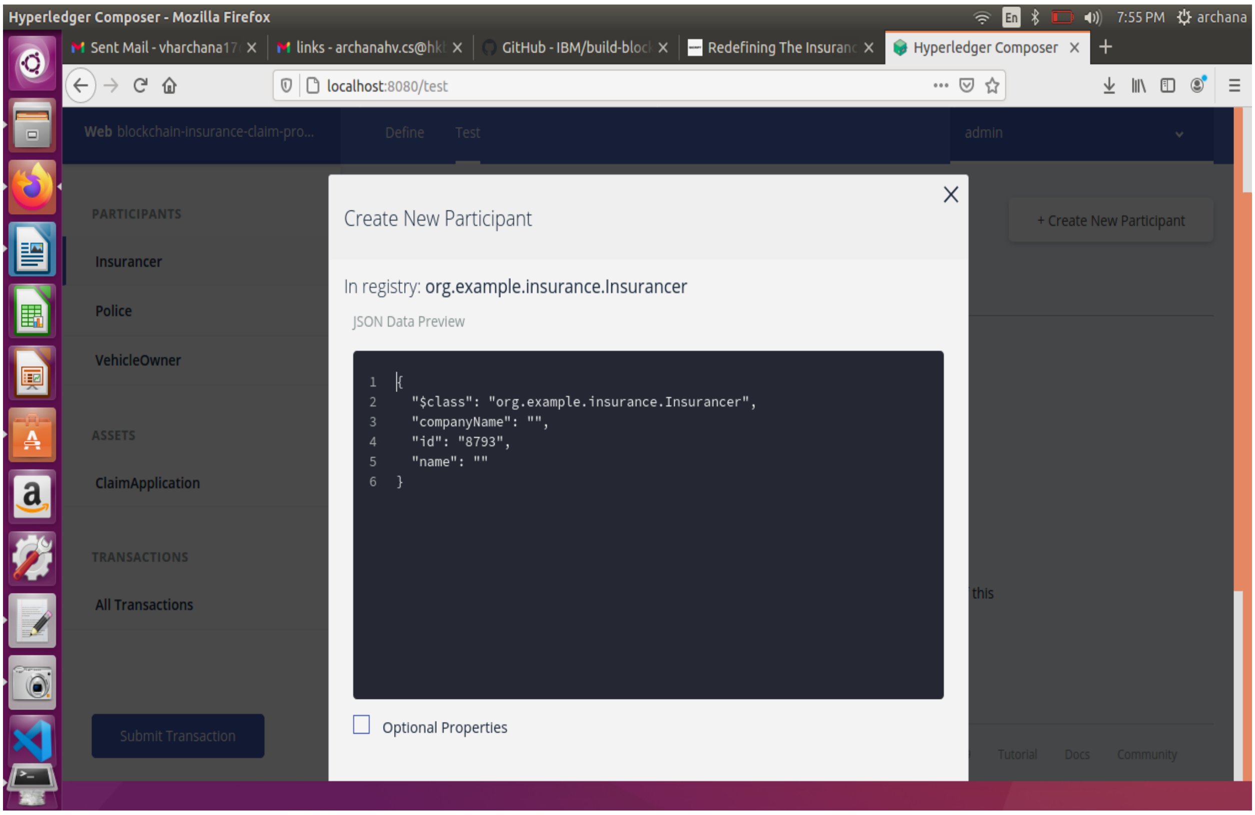
In registry: org.example.insurance.VehicleOwner

JSON Data Preview

```
{
  "$class": "org.example.insurance.VehicleOwner",
  "mailId": "",
  "phone": "",
  "id": "0540",
  "name": ""
}
```

□ Optional Properties

Figure 8: Screenshot of create vehicle owner participant.



Create New Participant

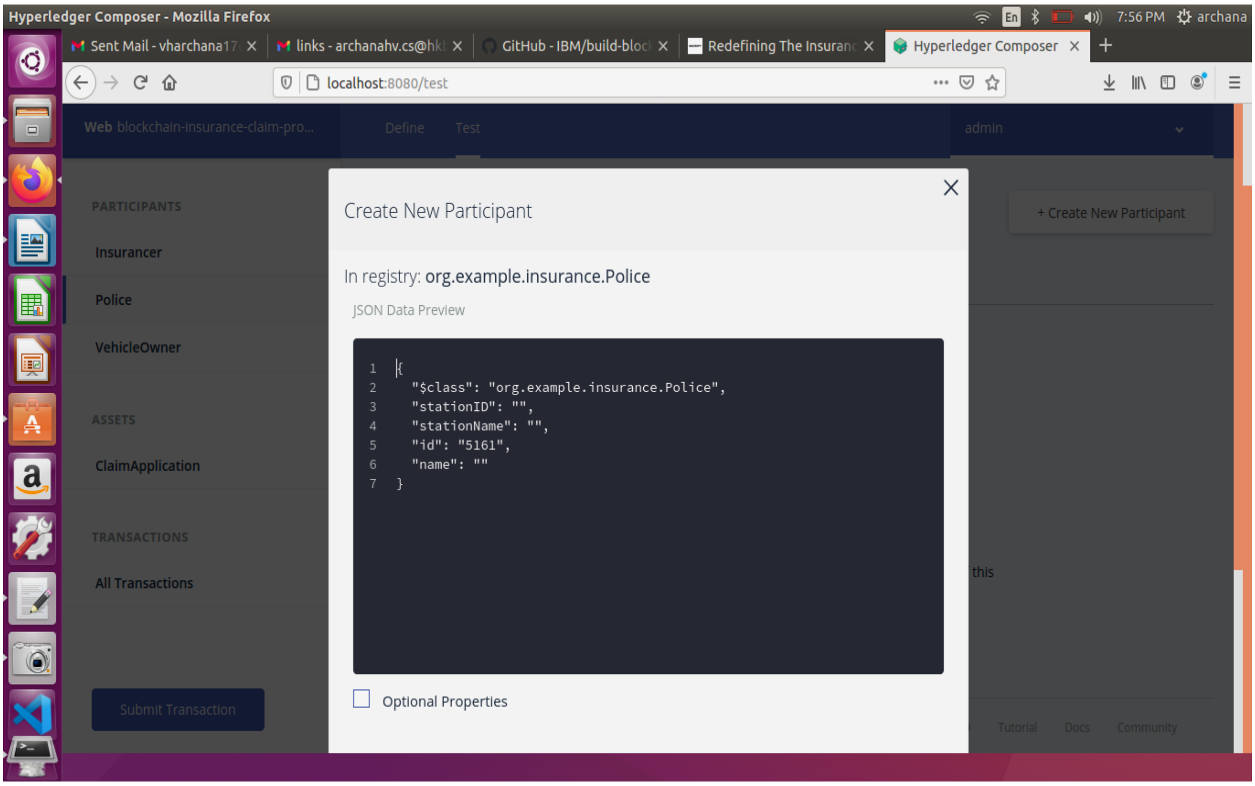
In registry: org.example.insurance.Insurancer

JSON Data Preview

```
{
  "$class": "org.example.insurance.Insurancer",
  "companyName": "",
  "id": "8793",
  "name": ""
}
```

□ Optional Properties

Figure 9: Screenshot of create insurance provider participant.



Create New Participant

In registry: org.example.insurance.Police

JSON Data Preview

```
{
  "$class": "org.example.insurance.Police",
  "stationID": "",
  "stationName": "",
  "id": "5161",
  "name": ""
}
```

□ Optional Properties

Figure 10: Screenshot of create police participant.

5 Results

Using hyperledger composer tools such as hyperledegr composer playground created business network page for insurance claim application processing.

6 Discussion

People's perceptions of blockchain technology have altered, and it has sparked a flurry of new business concepts. Plenty of insurance businesses have understood the importance of blockchain technology at this point. In the future, "blockchain + insurance" will be used at a higher level and in a bigger scope in the insurance industry. As more than just a finding of the research, the aforementioned goals were achieved:

- Smart contracts can be used to provide automatic claim settlement: From the occurrence of an insurance event through the reimbursement of payouts, all information and data will be created automatically across smart contracts, eliminating necessity investigation, hazard assessment, and evaluation.If a vehicle has insurance, for example, the record can be auto-generated and given to the insurance company, which will receive orders to pay the indemnity right away, which is better than prior insurance claims and reduces operational costs while improving customer service.
- By sharing information, you may verify a customer's identity security: Currently, the insurance sector faces the issue of workers or agents pressuring clients to accept surrender or survivor benefits. The main cause is that insurance companies do not have a system in place to control consumer identification. When a client receives a blockchain identity, the customer's identification information is no longer determined by the citizen ID but must be validated by all parties involved, minimizing the risk of numerous legal conflicts in the sector.
- Through entering data, you can develop a blacklist for the industry: Due to the insurance industry's low bar, a big percentage of agents breach the principle of good faith, and a large proportion of clients break laws and regulations. Practitioners and clients cannot be identified, and effective feedback cannot be delivered, because the industry lacks a blacklist platform. Blockchain data storage technology will be utilized to construct an industry blacklist as well as an open and transparent blacklist database to combat insurance fraud.
- Enhance the mutual insurance system by utilizing traceability technology. The inability of members to understand the flow of every fund is a major barrier restricting mutual insurance's expansion. Participants will have a clear awareness of each fund's spending and whereabouts thanks to the blockchain's information traceability technology,

allowing them to fully trust the mutual insurance organization. Mutual insurance groups will prosper over time if they operate in an atmosphere of complete trust.

- Defending against bogus claims by using the chain's subject information. Insurers typically lose authority of the true conditions of the insurance subject after establishing a property insurance contract. The complete process of tracking and managing the underlying assets of insurance is realized using blockchain technology to link the underlying assets to the chain, protecting the true contractual advantages and eliminating risks of repetitive insurance, target out of control, and false claims.

  In general, blockchain technology has shown a lot of promise in the insurance business. This will be especially essential in the ideological conflict and technological integration between blockchain technology and the insurance industry, as well as in "helping the real economy and avoiding financial risks."

7 Future study

The various ways in which the blockchain can transform the insurance industry:

- In FY2019, insurance crime in India was estimated to be around $45 billion. Insurer fraud, including data breaches, claim fraud, and qualifying fraud, is currently at an all-time level. The use of software created using smart contracts on the blockchain can minimize such fraudulent operations. This action will result in two things. With the use of blockchain technology, everyone will have access to information that cannot be changed.
- The decentralized technology that enables insurance firms to reduce costs and boost profits is strongly supported by the industry. The use of blockchain technology could assist save far more nearly $hundreds of millions of dollars annually in costs. Blockchain will eliminate data replication, enhancing design and validity while lowering loss or false claims. A further way that distributed ledger will cut expenses is by doing away with middlemen.
- Since then, everybody would be able to see information about the blockchain, service providers will find it simpler to communicate with one another, reducing mistake and fostering more trust between all parties. Because 3rd parties' validation is usually inadequate, the acquisition of new insurance contracts by various parties may result in misunderstanding and inconsistencies. However, access to information through a searchable, public digital network reduces waiting time and fosters trust amongst all parties. Both insurers and clients gain from open access to the real-time database since any modifications that result can be viewed and validated by all. It will be much less necessary for the insurer to rely on consultants for data, which will instantly streamline the insurance procedure.
- The peer-to-peer insurance system is a newer insurance model which is still in its early stages of development. It was created with the goal of enhancing transparency, minimizing threats, and decreasing malpractice. As a result of such a model's intricate structure, expansion concerns, and claim handling, insurance firms still encounter difficulties. Several researchers think that blockchain technology will soon help P2P insurance by minimizing fraud and enhancing scalability in order to address these problems.
- A blockchain transaction can assist insurance businesses in tracking meaningful results and, if needed, adjusting price, scheduling, or terminating. Launching innovative insurance policies, expanding their coverage, and locating customer groups in developing world and rural areas may all be done using the confirmed and confirmed data on the blockchain.

8 Conclusion

Distributed ledger technology is a potentially advanced technique for reliable online transactions, thus businesses whose operations are based on its use have a great potential. With the use of this platform, businesses engaging in multi-party companies can build responsibility based on dependable real-time information transmission. Which can provide useful and effective technologies that can be utilized in the development of systems, giving firms a competitive edge in terms of technology and operational efficiency. In this paper following contributions are made for the progress of vehicle insurance using blockchain technology. To begin with, problems related to online underwriting are solved successfully using blockchain technology. Secondly, this technology facilitates better oversight. Third, it is practical to avoid several claims for a single accident. Fourth, decentralized data access makes it easier to achieve data privacy and security. Fifth, providing time stamped transactions is beneficial in preventing fraud.

References

[1] International Association of Insurance Supervisors (IAIS). Available online: https://www.iaisweb.org/home (accessed on 20 August 2021).

[2] Raikwar, M.; Mazumdar, S.; Ruj, S.; Gupta, S.S.; Chattopadhyay, A.; Lam, K.Y. A Blockchain Framework for Insurance Processes. In Proceedings of the 2018 9th IFIP International Conference on New Technologies, Mobility and Security (NTMS), Paris, France, 26–28 February 2018. https://doi.org/10.1109/ntms.2018.8328731

[3] Limin, H.; Jianmin, Y. Application Research of Blockchain in the Field of Medical Insurance. In Proceedings of the 2019 3rd International Conference on Economics, Management Engineering and Education Technology (ICEMEET 2019), Suzhou, China, 18–19 May 2019.

[4] Zhang, X. Design and Implementation of Medical Insurance System Based on Blockchain Smart Contract Technology. Master’s Thesis, Huazhong University of Science & Technology, Wuhan, China, May 2019.

[5] Esposito, C.; De Santis, A.; Tortora, G.; Chang, H.; Choo, K.K.R. Blockchain: A panacea for healthcare cloud-based data security and privacy? IEEE Cloud Comput. 2018, 5, 31–37. https://doi.org/10.1109/mcc.2018.011791712

[6] Novo, O. Blockchain meets IoT: An architecture for scalable access management in IoT. IEEE Internet Things J. 2018, 5, 1184–1195. https://doi.org/10.1109/jiot.2018.2812239

[7] Wang, J.; Li, M.; He, Y.; Li, H.; Xiao, K.; Wang, C. A blockchain based privacy-preserving incentive mechanism in crowdsensing applications. IEEE Access 2018, 6, 17545–17556. https://doi.org/10.1109/access.2018.2805837

[8] Dorri, A.; Steger, M.; Kanhere, S.S.; Jurdak, R. Blockchain: A distributed solution to automotive security and privacy. IEEE Commun. Mag. 2017, 55, 119–125. https://doi.org/10.1109/mcom.2017.1700879

[9] Xia, Q.; Sifah, E.; Smahi, A.; Amofa, S.; Zhang, X. BBDS: Blockchain-Based data sharing for electronic medical records in cloud environments. Information 2017, 8, 44. https://doi.org/10.3390/info8020044

[10] Xu, J.; Xue, K.; Li, S.; Tian, H.; Hong, J.; Hong, P.; Yu, N. Healthchain: A blockchain-based privacy preserving scheme for large-scale health data. IEEE Internet Things J. 2019, 6, 8770–8781. https://doi.org/10.1109/jiot.2019.2923525

[11] Liu, X.; Wang, Z.; Jin, C.; Li, F.; Li, G. A Blockchain-based medical data sharing and protection scheme. IEEE Access 2019, 7,118943–118953. https://doi.org/10.1109/access.2019.2937685

[12] Kumar, S. (2020). Relevance of Buddhist Philosophy in Modern Management Theory. Psychology and Education, Vol. 58, no.2, pp. 2104–2111.

[13] Johari, R.; Kumar, V.; Gupta, K.; Vidyarthi, D.P. BLOSOM: BLOckchain technology for Security of Medical records. ICT Express 2021, in press. https://doi.org/10.1016/j.icte.2021.06.002

[14] S. Feng, Z. Xiong, D. Niyato, P. Wang, S. S. Wang, Y. Zhang, ―Cyber Risk Management with Risk Aware Cyber-Insurance in Blockchain Networks, ‖ 2018 IEEE Glob. Commun. Conf. GLOBECOM 2018 - Proc., 2018. https://doi.org/10.1109/glocom.2018.8648141

[15] Kumar, S. (2020). Relevance of Buddhist Philosophy in Modern Management Theory. Psychology and Education, Vol. 58, no.2, pp. 2104–2111.

[16] K. Panetta, ―5 trends emerge in the Gartner Hype Cycle for emerging technologies, ‖ Gartner. accessed December 8, 2020 unpublished.

[17] Blockchain.inf, ―Blockchain Charts‖ https://www. blockchain.com/ charts/, accessed September 17, 2020 unpublished.

[18] E. Kapsammer, B. Pröll, W. Retschitzegger, W. Schwinger, M. Weißenbek, & J. Schönböck, ―The Blockchain Muddle: A Bird's-Eye View on Blockchain Surveys‖, In Proc of the 20th Int Conf on Infor Integ and Web-based App & Ser (pp. 370-374). https://doi.org/10.1145/3282373.3282396

[19] K. Wang, & A. Safavi, ―Blockchain is empowering the future of insurance‖. Available at https://techcrunch.com/2016/10/29/blockchainis-empowering -the-future-of-insurance unpublished.

[20] F. Casino, T. K. Dasaklis, & C. Patsakis, ―A systematic literature review of blockchain-based applications: Current status, classification and open issues‖. Telemat Informatics 2019. https://doi.org/10.1016/j.tele.2018.11.006

[21] J. Mendling, I. Weber, W. V. Aalst, J. V. Brocke, C. Cabanillas, F. Daniel, S. Debois, C. D. Ciccio, M. Dumas, S. Dustdar, A. Gal, ―Blockchains for business process management-challenges and opportunities, ‖ ACM Trans. on Mgt Inf. Sys (TMIS). 2018 Feb 26;9(1):1-6. https://doi.org/10.1145/3183367

[22] S. Aggarwal, R. Chaudhary, G. S. Aujla, N. Kumar, K. K. Choo, A. Y. Zomaya, ―Blockchain for smart communities: Applications, challenges and opportunities, ‖ J of Net & Comp App. 2019 Oct 15;144:13-48. https://doi.org/10.1016/j.jnca.2019.06.018

[23] K. Yeow, A. Gani, R. W. Ahmad, J. J. Rodrigues, K. Ko, ―Decentralized consensus for edge-centric internet of things: A review, taxonomy, and research issues. IEEE Access. 2017 Dec 6; 6:1513-24. https://doi.org/10.1109/access.2017.2779263

[24] Wang, H.; Song, Y. Secure cloud-based EHR system using attribute-based cryptosystem and blockchain. J. Med. Syst. 2018, 42,152. https://doi.org/10.1007/s10916-018-0994-6

[25] Buterin, V. A next-generation smart contract and decentralized application platform. Ethereum White Paper 2014, 3, 36.

[26] Roy, S.; Das, A.K.; Chatterjee, S.; Kumar, N.; Chattopadhyay, S.; Rodrigues, J.J. Provably secure fine-grained data access control over multiple cloud servers in mobile cloud computing-based healthcare applications. IEEE Trans. Ind. Inform. 2018, 15, 457–468. https://doi.org/10.1109/tii.2018.2824815

[27] Wazid, M.; Das, A.K.; Kumari, S.; Li, X.; Wu, F. Provably secure biometric-based user authentication and key agreement scheme in cloud computing. Secur. Commun. Netw. 2016, 9, 4103–4119. https://doi.org/10.1002/sec.1591

[28] Sureshkumar, V.; Amin, R.; Vijaykumar, V.R.; Sekar, S.R. Robust secure communication protocol for smart healthcare system with FPGA implementation. Future Gener. Comput. Syst. 2019, 100, 938–951. https://doi.org/10.1016/j.future.2019.05.058

[29] Roy, S.; Chatterjee, S.; Das, A.K.; Chattopadhyay, S.; Kumari, S.; Jo, M. Chaotic map-based anonymous user authentication scheme with user biometrics and fuzzy extractor for crowdsourcing Internet of Things. IEEE Internet Things J. 2017, 5, 2884–2895. https://doi.org/10.1109/jiot.2017.2714179

[30] Banerjee, S.; Odelu, V.; Das, A.K.; Srinivas, J.; Kumar, N.; Chattopadhyay, S.; Choo, K.K.R. A provably secure and lightweight anonymous user authenticated session key exchange scheme for the Internet of Things deployment. IEEE Internet Things J. 2019,6, 8739–8752. https://doi.org/10.1109/jiot.2019.2923373

[31] Shuai, M.; Yu, N.; Wang, H.; Xiong, L. Anonymous authentication scheme for smart home environment with provable security. Comput. Secur. 2019, 86, 132–146. https://doi.org/10.1016/j.cose.2019.06.002

[32] Sehgal.P, Kumar.B, Sharma.M, Salameh A.A, Kumar.S, Asha.P (2022), Role of IoT In Transformation Of Marketing: A Quantitative Study Of Opportunities and Challenges, Webology, Vol. 18, no.3, pp 1-11.

[33] Yang, J.; Li, J.; Niu, Y. A hybrid solution for privacy preserving medical data sharing in the cloud environment. Future Gener.Comput. Syst. 2015, 43–44, 74–86. https://doi.org/10.1016/j.future.2014.06.004

[34] Soni, P.; Pal, A.K.; Islam, S.H. An improved three-factor authentication scheme for patient monitoring using WSN in remote health-care system. Comput. Methods Programs Biomed. 2019, 182, 105054. https://doi.org/10.1016/j.cmpb.2019.105054

[35] Masdari, M.; Ahmadzadeh, S. A survey and taxonomy of the authentication schemes in Telecare Medicine Information Systems.J. Netw. Comput. Appl. 2017, 87, 1–19. https://doi.org/10.1016/j.jnca.2017.03.003

[36] Amin, R.; Islam, S.H.; Biswas, G.P.; Khan, M.K.; Kumar, N. A robust and anonymous patient monitoring system using wireless medical sensor networks. Future Gener. Comput. Syst. 2018, 80, 483–495. https://doi.org/10.1016/j.future.2016.05.032

[37] Chen, L.; Lee, W.K.; Chang, C.C.; Choo, K.K.R.; Zhang, N. Blockchain based searchable encryption for electronic health recordsharing. Future Gener. Comput. Syst. 2019, 95, 420–429. https://doi.org/10.1016/j.future.2019.01.018

[38] Tanwar, S.; Parekh, K.; Evans, R. Blockchain-based electronic healthcare record system for healthcare 4.0 applications. J. Inf. Secur.Appl. 2020, 50, 102407. https://doi.org/10.1016/j.jisa.2019.102407

[39] Szabo, N. Smart contracts: Building blocks for digital markets. EXTROPY J. Transhum. Thought 1996, 18, 16.

[40] Szabo, N. The Idea of Smart Contracts. 1997. Available online: http://www.fon.hum.uva.nl/rob/Courses/InformationInSpeech/CDROM/Literature/LOTwinterschool2006/szabo.best.vwh.net/smart_contracts_idea.html(accessed on 20 August 2021).

[41] Vanstone, S. Responses to NIST's proposal. Commun. ACM 1992, 35, 50–52.

[42] Johnson, D.; Menezes, A.; Vanstone, S. The Elliptic Curve Digital Signature Algorithm (ECDSA). Int. J. Inf. Secur. 2001, 1, 36–63. https://doi.org/10.1007/s102070100002

[43] Burrows, M.; Abadi, M.; Needham, R. A logic of authentication. ACM Trans. Comput. Syst. 1990, 8, 18–36. https://doi.org/10.1145/77648.77649

[44] Sierra, J.M.; Hernández, J.C.; Alcaide, A.; Torres, J. Validating the Use of BAN LOGIC;

Springer: Berlin/Heidelberg, Germany, 2004; pp. 851–858. https://doi.org/10.1007/978-3-540-24707-4_98

[45] Hyperledger Fabric Docs. Available online: https://hyperledger-fabric.readthedocs.io/_/downloads/en/release-2.2/pdf/ (accessed on 20 August 2021).

[46] Foschini, L.; Gavagna, A.; Martuscelli, G.; Montanari, R. Hyperledger Fabric Blockchain: Chaincode Performance Analysis. In Proceedings of the ICC 2020—2020 IEEE International Conference on Communications (ICC), Dublin, Ireland, 7–11 June 2020; pp. 1–6. https://doi.org/10.1109/icc40277.2020.9149080

[47] Uddin, M. Blockchain Medledger: Hyperledger fabric enabled drug traceability system for counterfeit drugs in pharmaceutical industry. Int. J. Pharm. 2021, 597, 120235. https://doi.org/10.1016/j.ijpharm.2021.120235

[48] Marcus, M.J. 5G and IMT for 2020 and beyond. IEEE Wirel. Commun. 2015, 22, 2–3.

[49] D. E. Kouicem, A, Bouabdallah, H. Lakhlef, ―Internet of things security: A top-down survey, ‖ Computer Networks. 2018 Aug 4; 141:199-221. https://doi.org/10.1016/j.comnet.2018.03.012

[50] Q. E. Abbas, J. Sung-Bong, ―A survey of blockchain and its applications, ‖. In 2019 Int Conf on Art Intel in Inf and Comm (ICAIIC) 2019 Feb 11 (pp. 001-003). IEEE. https://doi.org/10.1109/icaiic.2019.8669067

[51] W. Liu, Q. Yu, , Li, Z., Li, Z., Su, Y. and Zhou, J., ―A BlockchainBased System for Anti-Fraud of Healthcare Insurance,‖ In 2019 IEEE 5th Int. Conf. on Compu. Commun. (ICCC) (pp. 1264-1268). IEEE. https://doi.org/10.1109/iccc47050.2019.9064274

[52] Ms. Elena Rosemaro. (2014). An Experimental Analysis of Dependency on Automation and Management Skills. International Journal of New Practices in Management and Engineering, 3(01), 01 - 06. https://doi.org/10.1109/iccc47050.2019.9064274

[53] H. Kim, M. Mehar, ―Blockchain in Commercial Insurance: Achieving and Learning Towards Insurance That Keeps Pace in a Digitally Transformed Business Landscape, ‖ SSRN Electron J 2019. https://doi.org/10.2139/ssrn.3423382

[54] M. Mainelli, B. Manson ―Chain reaction: How blockchain technology might transform wholesale insurance, ‖ How Blockchain Technology Might Transform Wholesale Insurance-Long Finance. 2016 Aug 1. Available at SSRN: https://ssrn.com/abstract=3676290.

[55] L. S. Howard, ―Blockchain insurance industry initiative B3i grows to 15 members, ‖ Insurance Journal. 2017; 6:2017.

[56] T. Q. Nguyen, A. K. Das, L. T. Tran, ―NEO Smart Contract for Drought-Based Insurance, ‖ 2019 IEEE Can. Conf. Electr. Comput. Eng. CCECE 2019, 2019. https://doi.org/10.2139/ssrn.3423382

[57] Y. Guo, Z. Qi, X. Xian, H. Wu, Z. Yang, J. Zhang, L. Wenyin, ―WISChain: An Online Insurance System based on Blockchain and DengLu1 for Web Identity Security, ‖ Proc. 2018 1st IEEE Int. Conf. Hot Information-Centric Networking, HotICN 2018, 2019. https://doi.org/10.1109/hoticn.2018.8606011

[58] Anokye Acheampong AMPONSAH1 *, Professor Adebayo Felix ADEKOYA2m, Benjamin Asubam WEYORI3Blockchain in Insurance: Exploratory Analysis of Prospects and Threats, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 12, No. 1, 2021. https://doi.org/10.14569/ijacsa.2021.0120153

[59] Mayank Raikwar∗, Subhra Mazumdar†, Sushmita Ruj†, Sourav Sen Gupta†, Anupam Chattopadhyay∗, and Kwok-Yan Lam∗ "A Blockchain Framework for Insurance Processes", 978-1-5386-3662-6/18/$31.00 ©2018 IEEE. https://doi.org/10.1109/ntms.2018.8328731

[60] D. Popovic, C. Avis, M. Byrne, C. Cheung, M. Donovan, Y. Flynn, C. Fothergill, Z. Hosseinzadeh, Z. Lim*, J. Shah, Understanding blockchain for insurance use cases A practical guide for the insurance industry. Presented at the Sessional Meeting of the Institute and Faculty of Actuaries [Staple Inn], 03 February 2020. https://doi.org/10.1109/ntms.2018.8328731

[61] Vukolic and Marko, "Rethinking permissioned blockchains" in ′ Proceedings of the ACM Workshop on Blockchain, Cryptocurrencies and Contracts, ser. BCC '17. New York, NY, USA: ACM, 2017. https://doi.org/10.1145/3055518.3055526

[62] C. D. Clack, V. A. Bakshi, and L. Braine, "Smart contract templates: essential requirements and design options", arXiv preprint arXiv:1612.04496, 2016.

[63] Ms. Nora ZilamRunera. (2014). Performance Analysis on Knowledge Management System on Project Management. International Journal

of New Practices in Management and Engineering, 3(02), 08 - 13. https://doi.org/10.17762/ijnpme.v3i02.28

[64] K. Christidis and M. Devetsikiotis, “Blockchains and smart contracts for the internet of things,” IEEE Access, vol. 4, pp. 2292–2303, 2016. https://doi.org/10.1109/access.2016.2566339

[65] I. Nath, “Data exchange platform to fight insurance fraud on blockchain,” in 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Dec 2016, pp. 821–825. https://doi.org/10.1109/access.2016.2566339

[66] W. Li, A. Sforzin, S. Fedorov, and G. O. Karame, “Towards scalable and private industrial blockchains,” in Proceedings of the ACM Workshop on Blockchain, Cryptocurrencies and Contracts. ACM, 2017, pp. 9–14. https://doi.org/10.1145/3055518.3055531

[67] H. Watanabe, S. Fujimura, A. Nakadaira, Y. Miyazaki, A. Akutsu, and J. Kishigami, “Blockchain contract: Securing a blockchain applied to smart contracts,” in Consumer Electronics (ICCE), 2016 IEEE International Conference on. IEEE, 2016, pp. 467–468. https://doi.org/10.1109/icce.2016.7430693

[68] Baliga, A. (2017). Understanding Blockchain Consensus Models. Retrieved October 2019, from https://www.persistent.com/wp-content/uploads/2018/02/wp-understanding-blockchainconsensus-models.

[69] Buterin, V. (2013, December). A Next-Generation Smart Contract and Decentralized Application Platform. Retrieved December 2019, from https://github.com/ethereum/wiki/wiki/WhitePaper.

[70] Aarti Patki1 · Vinod Sople2, “Indian banking sector: blockchain implementation, challenges and way forward”, Journal of Banking and Financial Technology, 11 May 2020. https://doi.org/10.1007/s42786-020-00019-w

[71] Mitt, Sven, “Blockchain Application - Case Study on Hyperledger Fabric”, 2018.

[72] Guy Zyskind, Oz Nathan, Alex ’Sandy’ Pentland “Decentralizing Privacy: Using Blockchain to Protect Personal Data”, 2015. https://doi.org/10.1109/spw.2015.27

[73] H. Sukhwani, J. M. Martnez, X. Chang, K. S. Trivedi, and A. Rindos, “Performance modeling of pbft consensus process for permissioned blockchain network (hyperledger fabric),” in 2017 IEEE 36th Symposium on Reliable Distributed Systems (SRDS), Sept 2017, pp. 253–255. https://doi.org/10.1109/srds.2017.36

[74] https://www.policybazaar.com/motor-insurance/general-info/articles/car-insurance-claim-process-guide/

[75] Blog written by Ashwin SoorajKudwa, Blockchain: Life and Vehicle Insurance, 08/23/18

[76] https://www.marutitech.com/artificial-intelligence-in-insurance/

[77] P. K. Sharma et al, “Software Defined Fog Node Based Distributed Blockchain Cloud Architecture”, IEEE Access, volume 6, February 1, 2018. https://doi.org/10.1109/access.2017.2757955

[78] Mrs. Monika Soni. (2015). Design and Analysis of Single Ended Low Noise Amplifier. International Journal of New Practices in Management and Engineering, 4(01), 01 - 06. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/33. https://doi.org/10.17762/ijnpme.v4i01.33

[79] L.Wang and R. Ranjan, ``Processing distributed Internet of Things data inclouds,'' IEEE Cloud Comput., vol. 2, no. 1, pp. 76_80, Jan. 2015. https://doi.org/10.17762/ijnpme.v4i01.33.

[80] M. Chiang and T. Zhang, ``Fog and IoT: An overview of research opportunities,'' IEEE Internet Things J., vol. 3, no. 6, pp. 854_864, Dec. 2016. https://doi.org/10.1109/jiot.2016.2584538

[81] Valentina Gatteschi et al., “Blockchain and Smart Contracts for Insurance: Is the Technology Mature Enough?”, Future Internet 2018, 10, 20. https://doi.org/10.3390/fi10020020

[82] Fran Casino et. al., “A systematic literature review of blockchain-based applications: Current status, classification and open issues”, 22 November 2018, 0736-5853/ © 2018 The Authors. Published by Elsevier Ltd. https://doi.org/10.1016/j.tele.2018.11.006

[83] Pradip kumarsharma et. al., “A Software Defined Fog Node Based Distributed Blockchain Cloud Architecture for IoT”, VOLUME 6, 2018. https://doi.org/10.1109/access.2017.2757955

[84] Jin Ho Park et. al, “Blockchain Security in Cloud Computing: Use Cases, Challenges, and Solutions”, 2017, 9, 164; doi:10.3390/sym9080164.

[85] Min Xu, Xingtong Chen and Gang Kou, “A systematic review of blockchain”, Xu et al. Financial Innovation (2019).

[86] Hany F. Atlam et.al., “Integration of Cloud Computing with Internet of Things: Challenges and Open Issues”, 2017 IEEE International Conference on Internet of Things. 2017 IEEE. https://doi.org/10.1109/ithings-greencom-cpscom-smartdata.2017.105

[87] Ms. Pooja Sahu. (2015). Automatic Speech Recognition in Mobile Customer Care Service. International Journal of New Practices in Management and Engineering, 4(01), 07 - 11. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/34. https://doi.org/10.17762/ijnpme.v4i01.34

[88] X. Sun, N. Ansari, and R. Wang, ``Optimizing resource utilization of a data center," IEEE Commun. Surveys Tuts., vol. 18, no. 4, pp. 2822_2846,4th Quart., 2016. https://doi.org/10.1109/comst.2016.2558203

[89] K. Valtanen, J. Backman, S. Yrjola, ―Blockchain-Powered Value Creation in the 5G and Smart Grid Use Cases, ‖ IEEE Access 2019. https://doi.org/10.1109/access.2019.2900514

[90] F. Z. Meskini, R. Aboulaich, ―A New Cooperative Insurance Based on Blockchain Technology: Six Simulations to Evaluate the Model, ‖ 2020 Int. Conf. Intell. Syst. Comput. Vision, ISCV 2020, 2020. https://doi.org/10.1109/iscv49265.2020.9204170

[91] A, Tapscott, D. Tapscott, ―How Blockchain Is Changing Finance, ‖ Harv Bus Rev 2017.

[92] O. I. Khalaf, G. M. Abdulsahib, H. D. Kasmaei, K. A. Ogudo, ―A new algorithm on application of blockchain technology in live stream video transmissions and telecommunications, ‖ Int J e-Collaboration 2020. https://doi.org/10.4018/ijec.2020010102

[93] Mr. Dharmesh Dhabliya, Ms. Ritika Dhabalia. (2014). Object Detection and Sorting using IoT. International Journal of New Practices in Management and Engineering, 3(04), 01 - 04. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/31. https://doi.org/10.17762/ijnpme.v3i04.31

[94] S. Olnes, J. Ubacht, M. Janssen, Blockchain in government: Benefits and implications of distributed ledger technology for information sharing. Gov Inf Q 2017. https://doi.org/10.1016/j.giq.2017.09.007

[95] P. K. Singh, R. Singh, G. Muchahary, M. Lahon, S. Nandi, ―A Blockchain-Based Approach for Usage Based Insurance and Incentive in ITS, ‖ IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON, 2019. https://doi.org/10.1109/tencon.2019.8929322

[96] Y. C. Chen, Y. P. Chou, Y. C. Chou, ―An image authentication scheme using Merkle tree mechanisms, ‖Futur Internet 2019. https://doi.org/10.3390/fi11070149

[97] L. S. Sankar, M. Sindhu, M. Sethumadhavan, ―Survey of consensus protocols on blockchain applications, ‖ 4th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2017, 2017. https://doi.org/10.1109/icaccs.2017.8014672

[98] S. Aggarwal, R. Chaudhary, G. S. Aujla, N. Kumar, K. K. Choo, A. Y. Zomaya, ―Blockchain for smart communities: Applications, challenges and opportunities, ‖ J of Net & Comp App. 2019 Oct 15;144:13-48. https://doi.org/10.1016/j.jnca.2019.06.018

[99] J. Lake, Understanding cryptography ‘s role in blockchains 2019. Unpublished.

[100]D. Puthal, N. Malik, S. P. Mohanty, E. Kougianos, G. Das, ―Everything You Wanted to Know about the Blockchain: Its Promise, Components, Processes, and Problems, ‖ IEEE Consum Electron Mag 2018. https://doi.org/10.1109/mce.2018.2816299

[101]J. J. Bambara, P. R. Allen, K. Iyer, R. Madsen, S. Lederer, M. Wuehler, Blockchain: A practical guide to developing business, law, and technology solutions. McGraw Hill Professional; 2018 Feb 16.

[102]T. K. Sharma, Public Vs. Private Blockchain: A Comprehensive Comparison 2019 unpublised.

[103]Mycryptopedia, 2018, Consortium Blockchain Explained, unpublished.

[104]C. Saraf, S. Sabadra, ―Blockchain platforms: A compendium, ‖ IEEE Int. Conf. Innov. Res. Dev. ICIRD 2018, 2018. https://doi.org/10.1109/icird.2018.8376323

[105]C. Christian, “Blockchain, cryptography, and consensus,” 2017.

Deep Learning-Based CNN Multi-Modal Camera Model Identification for Video Source Identification

Surjeet Singh, Vivek Kumar Sehgal
Department of Computer Science and Engineering and Information Technology,
Jaypee University of Information Technology, Waknaghat, Solan, Himachal Pradesh, India
E-mail: surjeetknmit@gmail.com, vivekseh@ieee.org

Keywords: convolutional neural networks, video forensics, audio forensics, camera model identification.

Received: September 12, 2022

Here is a high demand for multimedia forensics analysts to locate the original camera of photographs and videos that are being taken nowadays. There has been considerable progress in the technology of identifying the source of data, which has enabled conflict resolutions involving copyright infringements and identifying those responsible for serious offenses to be resolved. Video source identification is a challenging task nowadays due to easily available editing tools. This study focuses on the issue of identifying the camera model used to acquire video sequences used in this research that is, identifying the type of camera used to capture the video sequence under investigation. For this purpose, we created two distinct CNN-based camera model recognition techniques to be used in an innovative multi-modal setting. The proposed multi-modal methods combine audio and visual information in order to address the identification issue, which is superior to mono-modal methods which use only the visual or audio information from the investigated video to provide the identification information. According to legal standards of admissible evidence and criminal procedure, Forensic Science involves the application of science to the legal aspects of criminal and civil law, primarily during criminal investigations, in line with the standards of admissible evidence and criminal procedure in the law. It is responsible for collecting, preserving, and analyzing scientific evidence in the course of an investigation. It has become a critical part of criminology as a result of the rapid rise in crime rates over the last few decades. Our proposed methods were tested on a well-known dataset known as the Vision dataset, which contains about 2000 video sequences gathered from various devices of varying types. It is conducted experiments on social media platforms such as YouTube and WhatsApp as well as native videos directly obtained from their acquisition devices by the means of their acquisition devices. According to the results of the study, the multi-modal approaches suggest that they greatly outperform their mono-modal equivalents in addressing the challenge at hand, constituting an effective approach to address the challenge and offering the possibility of even more difficult circumstances in the future.

Povzetek: Razvita je metoda prepoznavanje izvornih kamer videoposnetkov s kombiniranjem zvočnih in vizualnih informacij z uporabo dveh DNN CNN tehnik.

1 Introduction

t should be noted that camera model identification has become increasingly important in multimedia forensic investigations, as digital multi-media content (including images, videos, audio sequences, etc.) is becoming more widespread and will continue to do so with the advancement of technology in the future. There is no doubt that a large part of this phenomenon can be attributed to the advent of the internet and social media, which have enabled a more rapid diffusion of digital content and, consequently, made it extremely challenging to trace their origins [28]. In forensic investigations, for instance, tracking the origins of digital content can be essential for identifying the perpetrators of such crimes as rape, drug trafficking, and acts of terrorism by tracing the origins of the digital content. There is also the possibility that certain private content may become viral through the internet, as has sadly happened in recent times with revenge porn, and there are other possibilities as well. It is therefore of fundamental importance to be able to retrieve the source of multimedia content in order to use it as a source [10]. The purpose of this paper is to determine the smartphone model used to acquire digital video sequences through the combined use of visual and audio information that has been extracted from the videos themselves. Due to the fact that there has been little work specifically done on identifying the video source in the forensic literature, we mainly focus on video source identification. In contrast, digital image analysis is one of the most commonly addressed aspects of digital imaging. Various peculiar traces left on the photograph when it was taken at the time when the image was taken can be used to identify the camera model that was used to acquire the image [3]. The two main approaches that can be used to identify the model of an image camera are defined as model-based and data-driven approaches in this vein. In contrast, the model-based approach, on the other hand, focuses specifically on exploiting the traces that are released as a result of the process of taking a digital image, in order to be able to identify the type of camera from the traces as a result of being able to identify the information through the process of tracing. A significant number of other processing operations and defects using the same kinds of picture acquisition pipeline, including dust particles left on the sensor and noise patterns [11], have been demonstrated to be able to convey information and provide accurate information about a camera model that has been employed. In the last few years, the advent of digital data and computational resources led to the development of data-driven approaches that far outperform the solutions based on models. The data-driven approach is able to capture the model traces instead of focusing on a specific trace left by the image acquisition process, as is typical in model-based methodologies since the interaction of various components allows the approaches to capture model traces as well. Data-driven methodologies that have been most successful are those based on learned features, which in other words are methods that feed digital images directly to a deep-learning paradigm in order to learn model-related features and to associate images with the original source data [32].

The Convolutional Neural Networks (CNNs) are now becoming the most popular solutions in this field. As far as our knowledge goes, the only study that explores the problem of camera model identification on video sequences has been published. In this paper, we use advanced deep-learning approaches to develop effective methods for identifying camera models using video sequences in order to identify small patches from video frames, which they then fuse into a single accurate classification result for each video. In this paper, we use advanced deep-learning approaches to develop effective methods for camera model identification using video sequences. Specifically, we are proposing a method for recognizing videos by automatically extracting suitable features from the visual and audio content of the videos by using CNNs that are capable of classifying them by combining these features. Using a mixed-modal approach to solve the identification problem, we define the proposed strategy as multi-modal since we extract visual and audio information from a query video to solve the problem. It is important to note that, for visual content, we use patches cropped from the frames and, for audio content, we use patches cropped from the Log-Mel Spectrogram (LMS) of the audio track in the video that is used to solve the identification problem light of this, the method suggested by falls into the mono-modal category, since the authors rely solely on the visual content of a query video in order to determine its classification. In order to identify multi-modal camera models, we propose two distinct approaches based on this information [25]. With both approaches, we make use of CNNs and feed them with a pair of visual and audio patches in order to feed them with information. Our first approach consists of comparing and combining the individual scores obtained from a pair of CNNs that have been trained following a mono-modal strategy, that is, one CNN has been trained to deal with only visual data and the other CNN has been trained to deal with audio data only. The second approach involves training a single multi-input CNN that can be used simultaneously to process both visual and audio patches at the same time. For each of the proposed approaches, we examine three different network configurations and data pre-processing, which are based upon effective CNN architectures that are well known in the state of the art of video processing in order to maximize the level of performance. We evaluated the results in relation to the Vision dataset, which comprises approximately 650 native video sequences along with their related social media versions, which amounts to almost 2000 videos recorded by 35 modern smartphones [15]. The videos on which we conduct the experiments are not only the original native ones; we also use the videos that have been compressed by the algorithms of WhatsApp and YouTube in order to explore the effects of data recompression as well as to investigate challenging situations where the training and testing datasets do not share similar characteristics. To provide a baseline strategy for comparing the achieved results, we also investigate the mono-modal attribution problems. There is no doubt that the vast majority of state-of-the-art works in multimedia forensics in recent years have always dealt with video sequences by either exploiting their visual or audio content in a separate manner or by both. It has only been recently that both visual and audio cues have been used for multimedia forensics purposes, but they do not address the task of identifying the camera model used in those works. It is proposed that we evaluate the results obtained by exploiting only visual or audio patches in order to classify the query video sequence in a mono-modal manner [29]. Based on the results of the experimental campaign which was conducted, it can be concluded that the multi-modal methodology proposed is more effective than mono-modal approaches. Accordingly, the pursued multi-modal approaches have shown to be significantly more effective than standard mono-modal approaches in terms of solving the problem in a more efficient way. Moreover, we find that data that undergo stronger compression (e.g., videos uploaded to the WhatsApp application) are more difficult to classify than data that undergo a weaker compression (e.g., files uploaded to YouTube) [20]. In spite of this, we found that multi-modal strategies outperformed mono-modal strategies also in this complicated scenario". For the purpose of extracting feature descriptors from a sequence of images and categorizing them according to their descriptors, the algorithm for categorizing videos uses feature extractors such as convolutional neural networks (CNNs), which are comparable to feature extractors used for image classification. Using deep learning-based video categorization, it is possible to examine, categories, and keep track of activities in visual data sources such as video streams by examining, categorizing, and tracking these activities. In addition to surveillance, anomaly detection, gesture recognition, and human activity recognition, video classification has many other applications as well.

1) For the purpose of classifying videos, the following steps can be taken as a guide to be taken as a guide.
2) Training materials should be created as part of the training process.
3) In order to classify videos, you need to select a classifier.
4) The classifier should be educated and assessed on a regular basis.
5) Using the classifier, you will be able to process the video data.
6) It is possible to train a classifier by using a large set of activity recognition video data sets, such as the Kinetics-400 Human Action Dataset, that are used for activity recognition.

A classifier can be trained by using a large-scale and high-quality set of activity recognition video data, such as the Kinetics-400 Human Action Dataset, which is a dataset collection composed of high-quality and large-scale activity recognition video data. Give tagged footage or video clips to the video classifier at the beginning of the process [39]. Using a deep learning video classifier that is composed of convolution neural networks, you may be able to forecast and categorize the videos based on the nature of the video input by using a deep learning video classifier that is constructed using deep learning techniques. As part of your process, you should ideally include evaluating your classifier as part of your analysis. It may also be possible to use the classifier to categorize activity based on a stream of live webcam video or a collection of video clips that are being streamed [17]. The Computer Vision Toolbox provides a variety of methods for training such as the slow and fast paths (Slow Fast), ResNet with (2+1) D convolutions, and two-stream Inflated-3D approaches as shown in Figure 1.

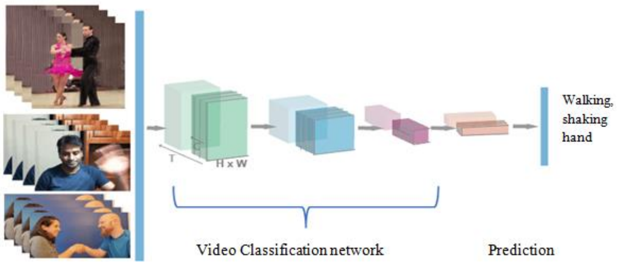


Figure 1: 3D techniques for training a classifier of video classification

1.1 An overview of camera calibration for DSLR cameras.

The manufacturers of DSLR cameras as well as other devices such as Canon, Nikon, and others often perform complex calibration algorithms before acquiring a scene image in their devices, which impacts the price of professional-level DSLR cameras considerably.

Therefore, it is necessary to develop effective, computationally less expensive, and affordable techniques of calibrating image-gathering equipment that are not inferior in quality to methods that are used abroad in order to make the process of calibrating equipment feasible and inexpensive for the masses [41]

1.2 Unique features of DSLR camera

There are a number of new digital forensic techniques that are being developed according to the unique characteristics of digital cameras that are closely related to the noise patterns from a few different kinds of DSLR cameras it is important to note that in order to solve the issues raised by the relevance of this work, it is necessary to limit one’s attention to those kinds of noise and distortion that can be observed and detected, i.e., those that can be determined technically (experimentally) through the measurement of noise and distortion parameters obtained or those that can be observed by expert observation and subjective evaluation [23]. There is the possibility that other types of noise that were overlooked can also be ignored as they have little impact on the final noise component in the image due to their small impact.

This research is structured as follows: In Section 2 we briefly mention a related topic called the background of videos, whereas in Section 3 we describe methods that can be used to identify videos as sources of information. Section 4 explains the method of forensic video analysis, Section 5 outlines the problem statement, Section 6 explains the research methods, and Section 7 explains the results of the study. This paper provides an evaluation of the resolution method to be used with the Kaggle dataset as part of the resolution scheme we propose for using the Kaggle dataset. During the analysis that has been conducted, the results that have been obtained along with the analysis that has been conducted will be discussed. In the end, some conclusions are reached based on the findings of the study

2 Related works

It is possible to identify the camera model used to capture the photos and video frames shown in this article by using the numerous odd traces that have been left on the images and video frames during the shooting process that have been captured. It is here that we will provide the reader with some background information about the typical acquisition process of digital photographs so that, in the future, the reader will be able to better understand. In the next step, we will take a look at how we define the Mel scale, as well as the audio content of video sequences, in the next step. The author points out that the LMS is an excellent tool for studying how an audio track has changed over time, as well as how its spectral content has changed over time [14]. The issue of identifying image camera models over the past few decades has been addressed in a variety of ways over the course of the past few decades [9][21][13][35]. It is the aim of these approaches to derive noise pattern characteristics for each camera model from the images or videos that are supplied to them. The noise patterns or traces in these cameras are believed to be a result of manufacturing defects and to be specific to each camera model [24].

2.1 Noise-based identification of digital video sources

In the field of multimedia forensics, there has been a great deal of attention paid to the task of blindly identifying the source device. By means of examining traces such as sensor dust and broken pixels, a number of strategies were put forth in order to identify the capturing device. When Lukas et al. first proposed the idea of utilizing Photo-Response Non-Uniformity (PRNU) noise to unambiguously define a camera sensor, they made a substantial advance in the understanding of the geometry of a camera sensor [7]. Because PRNU is a multiplicative noise, it cannot be effectively removed even by high-end equipment due to the fact that it is a multiplicative noise. The problem persists in the image even after JPEG compression at an average quality level has been applied to the quality level. In research on the viability of PRNU-based camera forensics for recovering photos from typical SMPs, it appears that alterations made to the photos by the user or the SMP could render the PRNU-based source identification useless

2.2 Analyzing the source of digitally identification videos

There is now digital identification technology built into new camera software to reduce the effects of unsteady hands on recorded footage caused by unsteady hands. In order to modify which pixels of the camcorder's image sensor are being utilized, this program evaluates the effect of user movement on which pixels on the image sensor are being utilized. It is generally true that image stabilization can be switched by the user on Android- based devices, but the camera software on iOS-based devices is not able to change this setting. In order to identify the source of videos shot with active digital identification using the PRNU fingerprint, the alignment of the fingerprints is disturbed during the identification process, which makes it impossible to identify the source of videos shot with active digital identification [30]. However, despite the fact that HSI has developed a reference side solution (which estimates the fingerprint from still photos), the problem still exists. Despite the fact that there are many variations in forensic video analysis techniques that could lead to the discovery of evidence, there are still many questions that need to be answered before they can be considered as being applicable. Additionally, forensic video analysis has shown to be more challenging than image analysis in terms of what it takes to make sense of the video's data. This is due to the fact that videos have more tightly compressed formats compared to picture formats [34]. An image frame is a series of images that make up the video that changes throughout time and evoke movement and change throughout time. A video is a video that contains a great deal of information that is encoded and decoded with the assistance of a mathematical technique called a codec, which encodes and decodes the information. In the multimedia file format, these previously encoded frames are wrapped up with tracks for the audio and metadata, as well as subtitles, and are known as multimedia files and are known as multimedia files.

3 Background

A number of strange traces were left on both the images and video frames that were captured during the shooting process. These traces have enabled researchers to determine the camera model that was used in order to capture these images and videos. We are trying to provide the reader with some background information about the typical digital picture collection pipeline in this section. In this way, they will be able to better comprehend the trace to which we refer in the next section. This will help them to understand it as well. After this, we define the Mel scale and the Log-Mel Spectrogram (LMS) of digital audio signals in order to be able to analyze the audio content of video sequences in the same way as we do the audio content of audio signals. LMS is a very valuable tool for examining the spectral and temporal evolution of an audio track. This is because it can be used to examine its spectral and temporal evolution based on its spectral and temporal characteristics.

3.1 A pipeline for acquiring digital images

In order to capture a picture with a digital camera or on a smartphone, we must initiate a complex process that involves numerous steps. This process involves numerous steps every time we use a digital camera. In a fraction of a second after pressing the shutter button, a short process begins which lasts only a fraction of a second. As soon as we are able to see the picture we have just taken, it stops. In general, the acquisition of a digital image does not follow a unique process.al image is not unique in most cases. There can be a significant variation in the vendors, the models of the devices, and the technologies that are onboard the devices. The picture acquisition pipeline can be thought of as a sequence of standard stages [42]. These are shown in Figure 2, which can be logically viewed as a sequence of standard steps.

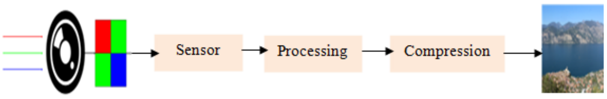


Figure 2: Acquisition of digital images.

3.2 A framework for analysis of forensic video

Compared to traditional photography-based evidence analysis in courts, forensic video analysis and the processing of multimedia evidence are still relatively novel fields compared to traditional photography-based evidence analysis in courts. It has become a growing trend over the last few years for a growing number of authoritative organizations, such as the Certified Forensic Video Analyst (CFVA) to recognize forensic video analysis as a significant objective norm, making its use in court more and more accepted. Forensic video analysis can be classified into the following four categories: Law enforcement forensic video analysis, forensic video and multimedia analysis, image/video comparison, and enhanced forensic video analysis. These are the major factors that are being focused on by the newest forensic video analysis techniques [4].In our work, we focus on ”enhanced forensic video analysis,” i.e. the analysis of video and data using the most advanced video analysis tools. This enhanced forensic video analysis architecture, shown in Figure 3, is comprised of three fundamental parts: crime scene analysis, data collection, video enhancement and analysis, and presentation and enlargement of the findings”.

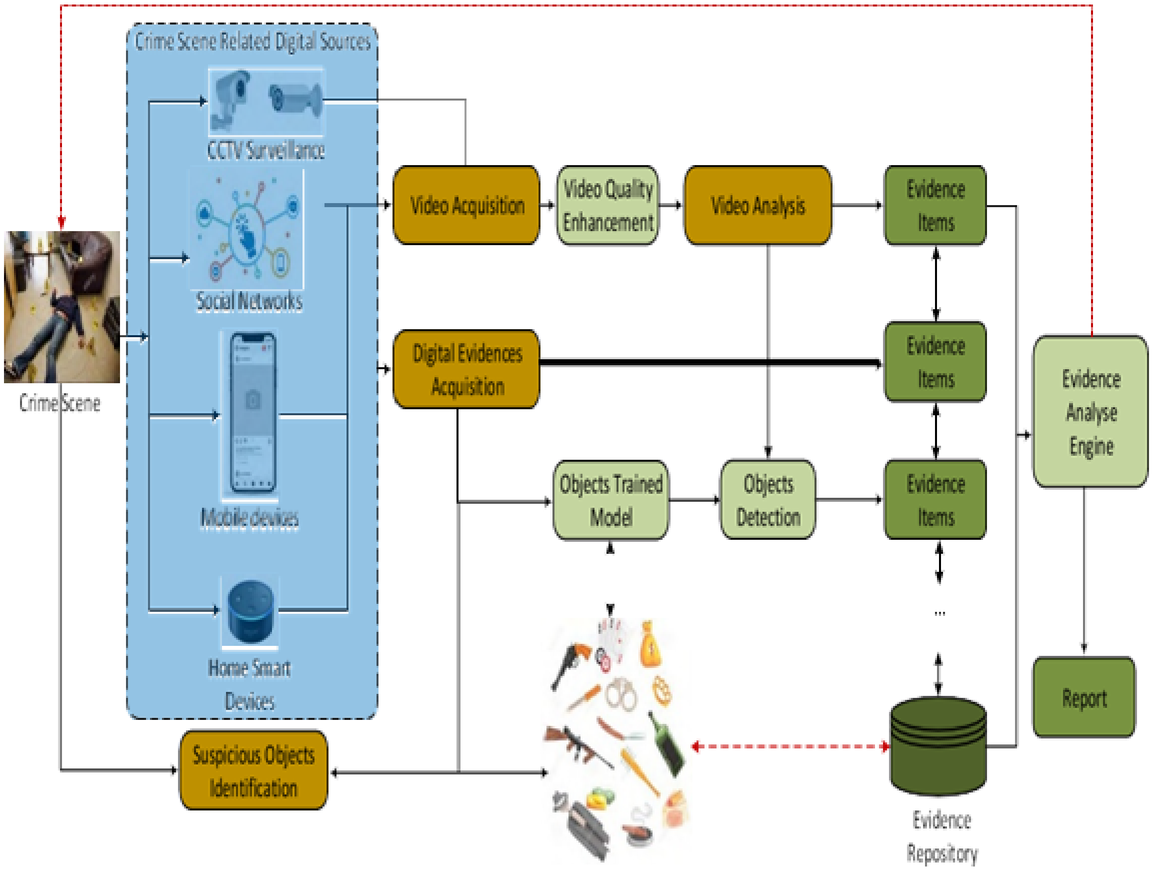


Figure 3: Enhanced forensic video analysis framework.

4 Method for the analysis of forensic videos

The preceding framework makes it obvious that there are two major categories of forensic video analysis that can be categorized in this manner an analysis of the content and type of video in a video. The retrieved pre-processed video is given to one or more CNNs in the CNN processing stage in order to extract unique characteristics among the many source camera models and categorize the original one[15].

4.1 A study of forensic video types and analysis

An obvious objective of forensic video analysis is to determine whether a video file has been unlawfully reproduced or tampered with. In addition, it is critical to determine whether the video has been altered in any way. It is also possible to identify concealed information in this research by identifying the video source and analyzing the video steganography to identify concealed information. In particular, the identification of the video source is a key evidence source [19]. This is because it determines whether the video source is a camera or a device that tokens the video or image as shown in Fig.4.It has been confirmed that forensic audio analysis, forensic video analysis, image analysis, and computer forensics are all distinct fields of study as determined by the American Society of Crime Laboratory Directors Laboratory Accreditation Board (ASCLD/LAB). A large number of private, public, and state/local law enforcement organizations are now creating digital and multi-media sections within their organizations that may cover some or all of these disciplines. There are some agencies where the same person may conduct examinations for different agencies. It is quite common for examiners in large agencies, at the federal and state levels, and in one field to specialize after years of training to become subject matter experts in their area. There are a number of ways in which video evidence can be enhanced [40]. It is very critical to submit the highest quality video recording in order to receive the most effective results from the enhancement process. A digital file or analogue copy that has been compressed with extra compression, if sent in for examination, may not be able to undergo the enhancement process. This is because it has been compressed with extra compression.

4.2 Enhancement of videos techniques

In order to achieve this goal, a wide variety of approaches has been used over the past decade to improve the quality of video. Several of these approaches have been developed for video monitoring systems intelligent highway systems, safety-monitoring systems, and a variety of other applications. As an example, [36]. have developed a method for identifying luggage from low-quality video footage by incorporating color information into the video footage. In order to identify the moving direction of an object, human-like temporal templates can be constructed and aligned with the appropriate parameters in order to identify the direction in which the object is moving. A number of authors have suggested that a system for detecting luggage should be created. As stated in Chuang et al., the purpose of the study was to detect missing colors using a ratio histogram. This variable is the ratio of the color histograms [31]. To find the missing colors, a tracking model should be used. From low-quality videos, forensics' primary goal is to extract as much information as possible from them in order to assist in the investigation process. It is the purpose of this section to present strategies for improving videos so that more information can be obtained from them. In low-quality videos/images, the likelihood of detecting additional information can be significantly enhanced using histogram equalization (HE)-based approaches compared to conventional approaches. Here is an example of how a webcam can be used to recognize objects using the suggested technique shown in Figure 4.

5 Problem formulation

In the present paper, we focus on the problem of identifying camera models from video sequences based on video content. As a primary focus of our research, we plan on identifying the source camera model from digital video sequences [33]. This has been attributed to the fact that digital image analysis has been extensively investigated in the forensic literature, without- standing results. In this study, we specifically work with video sequences that have been captured from a variety of smartphone models. This paper describes a novel method for combining informational and auditory information of videos under con- sideration to provide a comprehensive analysis of the videos under consideration [8]. We will first look at the classic mono-modal issue that seeks to identify the source camera model of a video sequence based on only visual or aural information, which will be discussed in the following sections. Next, we present the actual multi-modal problem identified in this research, which uses both visual and aural cues to identify the source of the sound.

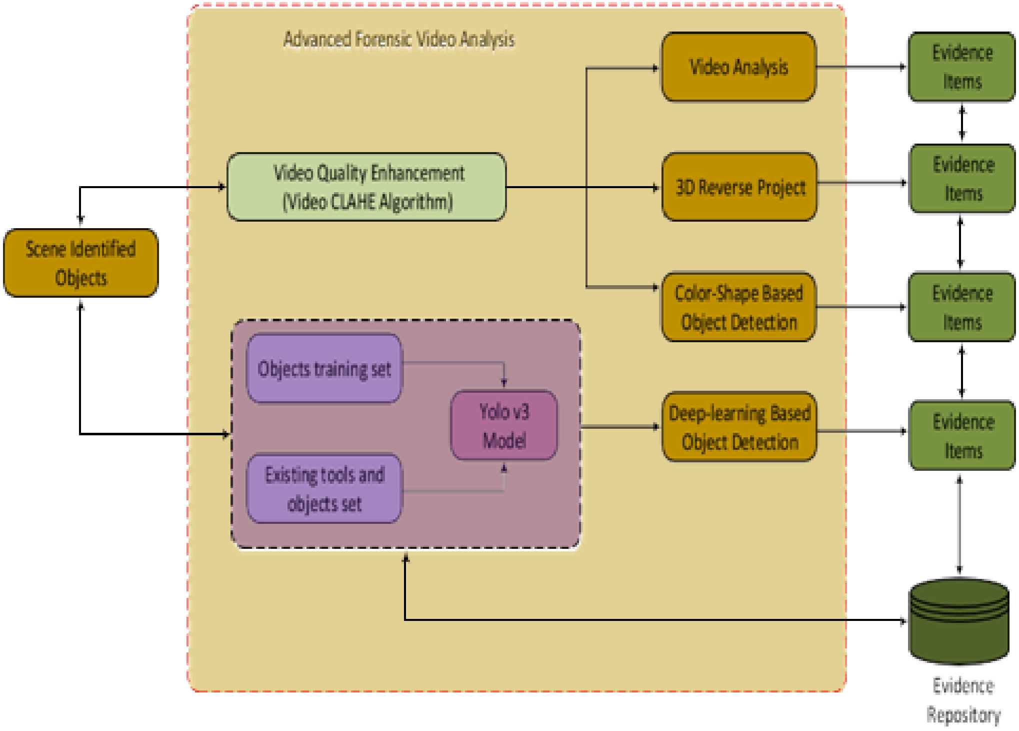

Advanced Forensic Video Analysis
Scene Identified Objects
Video Quality Enhancement (Video CLAHE Algorithm)
Objects training set
Yolo v3 Model
Existing tools and objects set
Video Analysis
3D Reverse Project
Color-Shape Based Object Detection
Deep-learning Based Object Detection
Evidence Items
Evidence Items
Evidence Items
Evidence Items
Evidence Repository

Figure 4: Video analysis procedures for advanced forensics.

5.1 Mono-Modal camera model identification

As a result, the problem is identified in the form of the device model, which was designed to acquire a particular media type in a single modality. When, for instance, an image has been captured, it is useful to know the model of the camera that was used to capture it. This is so that we can trace it back to its origin. In addition, if you have an audio recording, please include the model of the recorder that was used, along with the recording [26]. According to the mono-modal model attribution, in the context of a video, which is the situation we're interested in, the attribution of the device type that shot the video is identified solely based on the visual or auditory information contained within it.

5.2 Multi-Modal camera model identification

In the case of a video sequence, the challenge of multi-modal camera model identification is reduced to identifying the model of the device that recorded the video, taking both visual and aural information from the video sequence as input. In this example, we will consider a closed-set identification process that involves determining the camera model used to shoot a video sequence from a set of known devices that have been utilized in the past [38]. Assuming that the video being studied has been captured using a device from a device family familiar to the investigator, the investigator will assume that the video has been captured with a device of that device family. There is a possibility that the investigator will incorrectly assign a video to one of those devices if it does not originate from one of those devices.

6 Methodology

In this study, we present a method for identifying closed-set multi-modal camera models on video sequences that can be applauded in further research. In Figure 5 shows the main scheme of the proposal approach. Based on the visual and aural content of the video under consideration, we can determine the type of smartphone model that was used to capture the video. Using visual and auditory cues extracted from query video sequences, we input them into one or more CNNs that are capable of detecting the differences between different camera models used in the source video cameras based on their visual and auditory cues [2]. Two major steps comprise the proposed strategy, briefly:

1) Preprocessing and content extraction: The extraction of visual and auditory information from the videos under investigation, as well as the manipulation of the data before it is fed to CNNs, is referred to as pre-processing and content extraction.
2) There is a CNN processing block that consists of an extraction block that parses text into features and a classification block that consists of a CNN.

6.1 Content extraction and pre-processing

As part of the extraction and pre-processing step, visual and audio content is altered, as well as data standardization is performed.

There are three phases in this approach shown in Figure 6 that are involved in the extraction and pre-processing of visual content from the movie under analysis. These are:

1) It is possible to extract color frames from Nv that are equally distant in time and are spread out over a long period of time [12]. There are two sizes of video frames, which are Hv and Wv, and their sizes are determined by the resolution of the video being analyzed.
2) It is a raBy means of a random process, NPvcolour patches of the size HPV WPV are extracted at randomly to feed data into CNNs, patch normalization is carried out to ensure there is zero mean and unit variance.

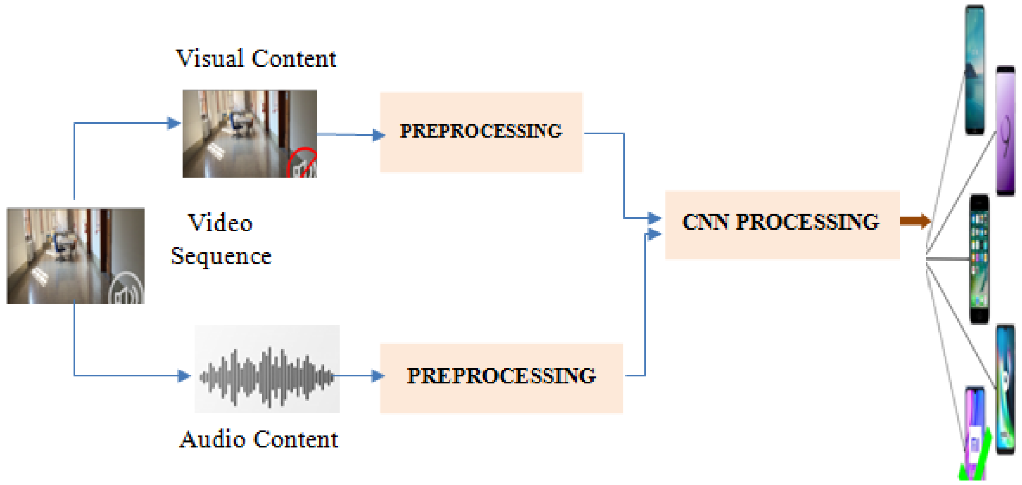


Figure 5: Pipeline of the proposed method.

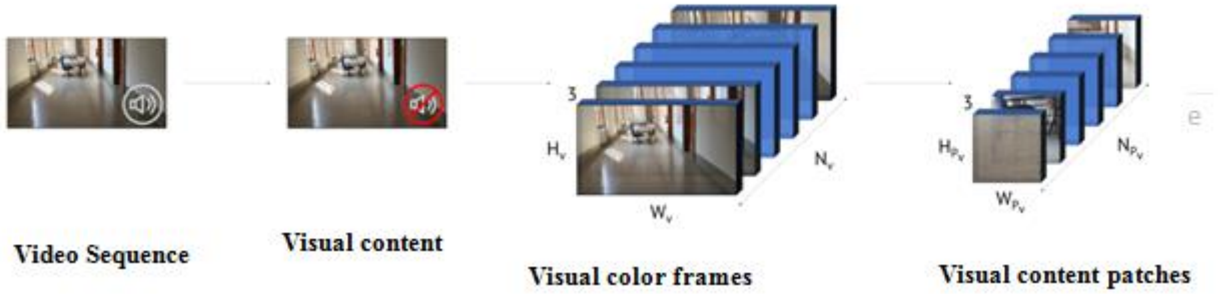


Figure 6: Process of creating a visual patch from a video stream.

There are three steps involved in the extraction and preparation of the audio material of the movie under examination shown Fig.7.

1) An extraction of audio content from the LMS L linked to the video sequence is performed. Considering this, it is clear that the LMS is a very useful tool for audio data and has been employed as a valuable feature for audio and speech classification and processing in a number of different studies. A number of audio characteristics were extracted from the magnitude and phase of the signal STFT during some exploratory experiments and it was determined that the LMS (based on the magnitude of the STFT signal) provided the best results. In the case of phase-based methods [1], LMS achieved an accuracy rate of less than 80%. As shown in the image below, the LMS L is a matrix of dimension Ha Wa, in which rows represent temporal information (which varies in length with the length of the video) and columns represent frequency content in Mel units.
2) Extraction of NPa patches of size HPaWPa randomly from L at random.
3) In order to achieve zero mean and unitary variance, patch normalization has been employed, as previously explained as described for the visual patches.

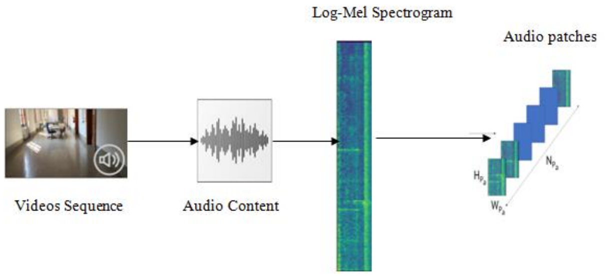


Figure 7: Audio patches extraction from a video sequence.

6.2 CNN processing

When the pre-processed information is retrieved, it is given to one or more CNNs in the CNN processing stage in order to extract distinct features based on the many source camera models and classify them accordingly demonstrate how it is possible to solve the mono-modal camera identification problem by feeding the retrieved visual or auditory data to a CNN [18]. In principle, any CNN architecture that is capable of classifying data could be employed at this stage; however, we discuss our choice in more detail in the next section.il in the next section. The final layer of the classification network is a fully connected layer with a number of nodes equal to the total number of models, M, where each node corre- sponds to a particular model of camera in the network. In this case, we are planning to produce an M-element vector with the name y, in which each element ym represents the likelihood that the model associated with the node was able to obtain input data. The node was able to obtain input data. We can extract it from the classification process by selecting the anticipated model m.

6.3 Early fusion methodology

As in the first method, the second method, called Early Fusion, involves combining two CNNs together to create a CNN with multiple inputs. In order to form the union, the final fully-connected layers of the two networks are concatenated, and three fully-connected layers are added until the prediction is formed As a result, the camera type is determined by the layer's dimensionality shown in Fig.8.

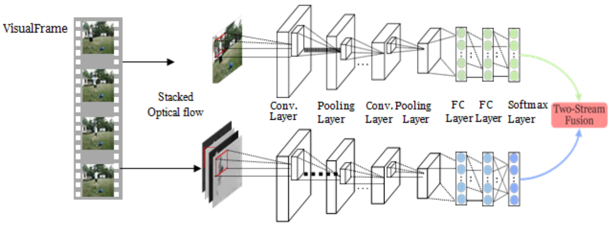


Figure 8: Early Fusion method pipeline.

Using the visual and audio patch pair, each Early Fusion forecasts the estimated camera model based on its estimated camera model in the final fully connected layer, yEF is the score obtained as a result of the final fully connected layer [37]. In the training phase, we use visual and audio patch pairs as a means of training the entire network. It is important to note that this is not the case with Late Fusion, since there is no separate training for the visual and audio branches. Similarly, both the training and testing phases are similar to those of the monomodal technique, except that we are distributing visual and audio patch pairs to the entire network this time instead of single patches (e.g., limited to visual or audio content). As shown in Figure8., the Early Fusion method's workflow is depicted in a flow chart. The size of the fully-connected layers' input and output features are also provided in order to facilitate the design [16]. In addition, it is worthwhile to mention that the output feature at the final layer of the network has a size equal to M, which is the number of camera models that have been evaluated

6.4 CNN architectures

A CNN called EfficientNetB0 and a CNN called VGGish are the two CNNs we are using in order to solve this problem.

EfficientNetB0 is a member of the recently proposed Efficient Net family of CNN models. It has demonstrated excellent performance in multimedia forensics tasks and is one of the most promising models within the family We chose this Efficient Net model as it is the most basic model that we could use. As a result, we have a lot more time to experiment with different evaluation configurations as it enables faster training phases. It has also been demonstrated, also through preliminary experiments, that there is no evidence of a significant change if one uses parameters like This is an evaluation of EfficientNetB0's performance when compared to computationally heavier network models with more parameters that require more computation [6]. There are a number of CNNs being used for audio classification, including the VGGish CNN, which has been inspired by the well-known VGG networks used in image classification. In order to solve this problem, we are employing two CNNs, one referred to as EfficientNetB0 and the other referred to as VGGish. In the recently proposed Efficient Net family of CNN models, EfficientNetB0 is one of the members of the Efficient Net model family. Among the highest performing models within the family, it has demonstrated excellent performance in multimedia forensics tasks, and is one of the most promising models within the family [27]. We chose the Efficient Net model because it is the most basic model, we can apply to achieve our goals.ls. Therefore, we have a lot more time to experiment with different evaluation configurations. This is because have a much faster training phase due to the fact that we have more time to play around. As we have already seen through preliminary experiments, it has also been demonstrated that there is no evidence of a significant difference if one uses parameters like this is an evaluation of EfficientNetB0's performance when compared to computationally heavier models with more parameters that require more computation than EfficientNetB0 [5].

A number of CNNs are being used for audio classification, including the VGGish CNN, which is based on the well-known VGG network that is used for image classification, that has been inspired by the well-known CNNs that are used for audio classification. After exploring the dataset, you need to create the training set and the validation set. The training set will be used to train the model, while the validation set will be used to assess the model that has been trained. It is recommended to ex- tract frames from each video that is part of the training set and the validation set. After preprocessing these frames, train a model on the training set of frames after the preprocessed frames have been used. For the purpose of evaluating the model, use the frames from the validation set as input. In the case that the performance on the validation set is satisfactory, we can use the trained model to categorize additional videos. According to Figure 9, the top portion of the figure shows the flow of the spatial stream's processing data. The CNN used for categorizing pictures is built in a similar way to a conventional deep CNN used for image categorization. In this method, each video frame is used as the input to the network, and then on top of that are added a number of convolutional layers, pooling layers, and fully connected (FC) layers.

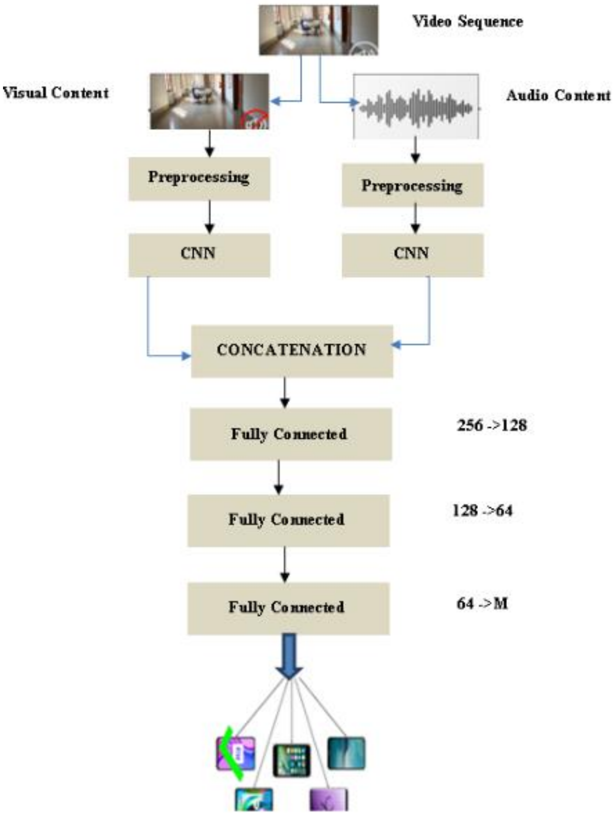


Figure 9: Processing pipeline for CNN's two-stream feature extraction.

7 Results

In this section, the dataset is processes first for experimental setup (i.e., the network training parameters and the configurations that we use in order to train the network). It is then reported what the evaluation metrics were, along with comments on what the results achieved.

7.1 Dataset

This study uses video sequences that are part of the Vision dataset. This is a recently released picture and video collection that has been created specifically for multimedia forensics investigations. Approximately 650 native video sequences were captured by 35 current smartphones and tablets, as well as their social media counterparts, as part of the Vision dataset. There are around 2000 video sequences in the collection, each of which has a clear indication of the source device from which it was captured. In our trials, we selected non-flat movies (that is, movies displaying natural situations with objects) both from the original source (that is, videos that are obtained through the camera on a smartphone without any post-processing) and those that have been compressed by WhatsApp and YouTube.

In order to achieve the granularity, we seek in our analysis, we aggregate movies from different devices that belong to the same model. This allows us to analyze them at the model level. The videos taken from the device D04, D12, and D17 As per the Vision dataset nomenclature provided in this publication, lines D21 and D22 have been omitted because they cause problems with the extraction of frames or audio tracks. In addition, we exclude original videos that are not available on WhatsApp or YouTube in a compressed form WhatsApp or YouTube.

Unlike most other video analysis services out there, we don’t just focus on high- resolution videos: while the majority of native videos have resolutions equal to or greater than 720p, we also examine native sequences with resolutions as low as 640480. As a result, we have 1110 videos that are around one minute in length, which were captured by 25 different cameras. In order to test the classification performance of the suggested technique, we use the available information about the model of the source camera as the ground truth for each video sequence. We extract 50 frames from each video sequence, equally distant in time and dispersed throughout the entire duration of the video sequence, in order to obtain the visual content. As a result, we extract 10 patches per frame (taken in random positions) for a total of NPv = 500 color patches per video. With 256 256 pixels as the patch size, we are able to achieve good results. Kaggle’s dataset with ten classes and 275 instances may have been used as the basis for the feature extraction process. This could have resulted in issues such as overfitting and a decrease in the accuracy of the prediction. This was the reason why we constructed a fresh dataset with 1300 cases from three classes in order to overcome these situations: iPhone 6s, Xiaomi Note 4x, and Samsung Galaxy J7. Our next step was to introduce two new classes into the system. There are 275 Samsung Galaxy Note 3 and HTC One M7 examples included in the Kaggle dataset is shown in Table 1. In order to extract the features of the proposed model, the dataset was given to the model and the features were extracted to the model and the features were extracted. We categorized the camera models based on the characteristics retrieved from the retrieved data.

Table 1: Details of the dataset.

Model Name	**Number of Instances**	**Acquired From**
IPhone 6s	1500	self
Xiaomi Note 4x	1560	self
Samsung Galaxy j7	1600	self
Samsung Galaxy Note 3	1000	Kaggle
HTC One M7	550	Kaggle

According to Table 2, we present the error rate and the average confidence score for the test split of the patch dataset for different values of which have been found to lead to high misclassifications of adversarial instances while FGSM has not resulted in meaningful visual changes for untargeted attacks. Based on the patch test split, we discover that using = 0.005 provides the best compromise between error rate and apparent changes in the image, with the result that the trained DenseNet model detector has an average error rate of 93.1 percent and an average confidence level of 95.3 percent. When the value of increases, it should be noted that the manipulations become more visible as the value of rises.

This table displays our trained DenseNet model’s error rate and confidence score following an untargeted FGSM assault to the test split. The second experiment, which is the CFA interpolation, is performed by simply taking the second set of features alone, which is the second set of features. According to the last analysis, the accuracy of the result was 86.93%. It is considered acceptable, but not enough, and it is still less than the result of the first experiment of co-occurrences alone, which was considered acceptable, but not enough. In order to achieve 97.81% accuracy on average, we combined the two feature sets into one and implemented them together. The average score achieved by all three sets was 98.75%.

Table 2: DenseNet model’s error rate and confidence score.

Value	Error Rate (%)	Confidence Score (%)
0.01	97.3	97.8
0.02	94.8	91.0
0.03	92.6	93.9
0.04	93.7	92.8
0.05	98.4	94.8
0.06	96.7	98.6
0.07	91.5	99.4
0.08	90.6	97.1
0.09	92.0	92.0
0.11	91.4	91.2

According to Table 3, all the experiments mentioned above along with their accuracy rates are shown. The results of these experiments are presented in Table 3. The table below displays both the overall test accuracy as well as the test accuracy for each ConvNet for each of the three settings (flat, indoor, and outdoor) and each of the three compression types (native (NA), WhatsApp (WA), and YouTube (YT). Furthermore, these results are in agreement with tests that were conducted using N I-frames per movie for both training and testing. On the basis of PRNU, the best accuracy in the trials exceeds that of the limited counterparts by a large margin in each of the scenarios and compression types that were tested. On the VISION data set. As a comparison, we also conducted

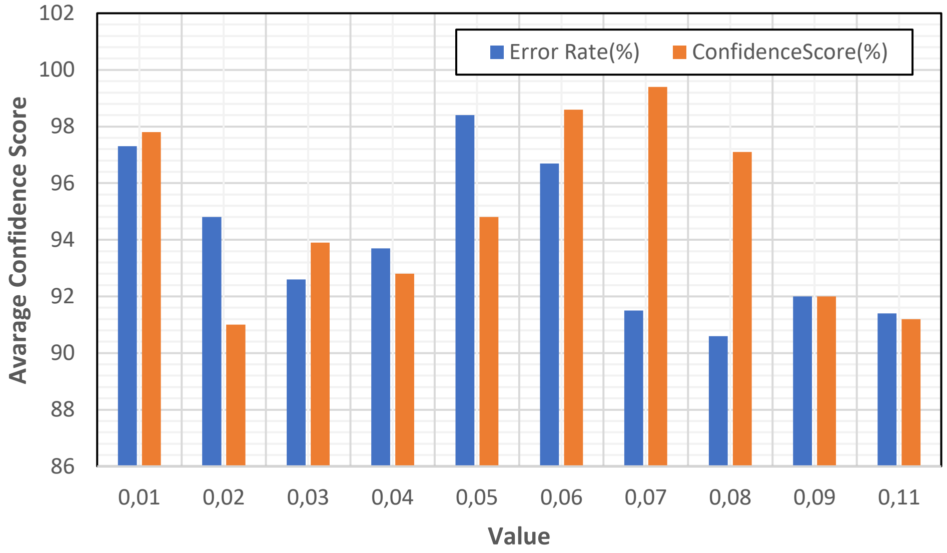


Figure 10: Comparison of the proposed method with other methods.

Table 3: Classification accuracy based on VISION data.

Model	N	Constraint Type	Overall	Flat	Indoor	Outdoor	WA	YT	NA
ResNet50	60	Conv	55.20	64.81	50.74	41.71	55.10	51.60	62.80
ResNet50	60	Conv	55.20	64.81	50.74	41.71	55.10	51.60	62.80
MobileNet	60	None	71.57	85.32	62.87	75.45	78.66	67.96	71.66
MobileNet	60	Conv	56.18	64.74	47.21	56.51	53.60	46.20	53.00
MobileNet	60	PRNU	62.70	63.96	53.11	61.12	58.80	63.50	67.30
MobileNet	60	None	75.87	76.92	64.62	75.02	74.84	77.68	75.90
MobileNet	60	PRNU	61.74	65.96	54.14	67.14	57.81	65.54	68.31

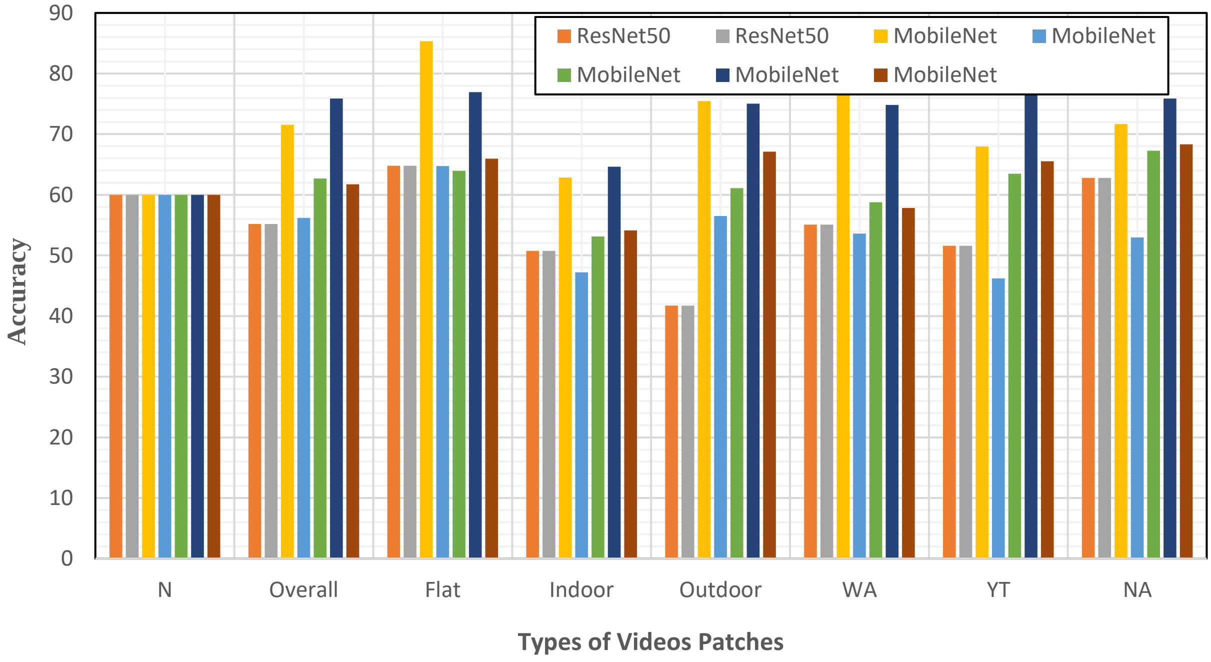


Figure 11: Classification accuracy of camera for proposed method.

Table 4: Compares the accuracy of MobileNet when it is compared to different counts of I-frames per video (I-fpv).

I-fpv	Overall	Flat	Indoor	Outdoor
1	69.12	71.1	57.5	76.5
5	72.31	79.8	59.6	75.4
30	74.10	82.1	62.3	76.0
50	73.51	81.5	61.6	75.4
100	73.71	82.1	61.6	75.4

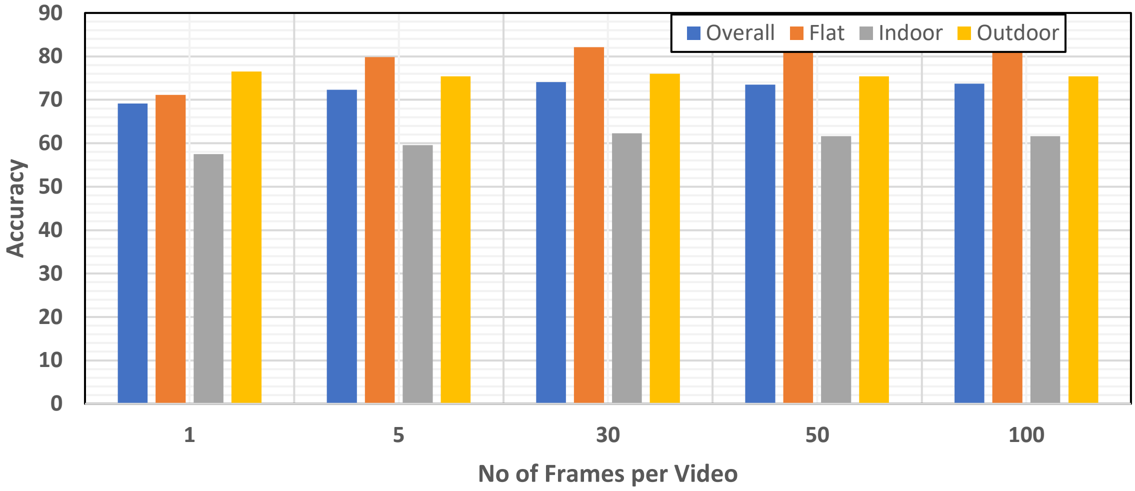


Figure 12: Test accuracy of mobile, net frames per videos.

the same experiment using the I-frames and the results are shown in Table 4. The results of the study show that the model achieves a high level of accuracy even when only a small number of tests I-frames are used. In addition, due to the short length of the movies included in the VISION data set, there are fewer I-frames available. Thus, even though we try to extract more I-frames, our accuracy remains the same, despite extracting more I-frames. As a result of our experience, we believe that the most effective overall strategy would be to apply the Late Fusion methodology in conjunction with configuring the EE192 according to our experience. With regard to native video sequences as well as YouTube video sequences, it consistently reports the most accurate results, regardless of whether it is a cross-test or not, and regardless of whether the test is a non-cross test or a cross-test. It is interesting to note that the cross-test results, including WhatsApp data, are on par with those of the other two configurations, if not a bit below. As a result of the fact that the trained CNNs. in this configuration are very adaptable to the data that they are shown during the training phase (i.e., patches selected from native or YouTube video sequences), they become less general and highly sensitive to significant data compression, such as that applied by WhatsApp, explaining the poor performance.

8 Conclusions and future works

The outcomes demonstrate that the suggested multi-modal methods are much more productive than traditional mono-modal methods. This research proposes a brand-new multi-modal methodology for identifying closed-set cameras models from digital video sequences that can be applied to digital video sequences. The overall objective of this research is to identify the smartphone model used to capture a query video by using visual as well as audio data from the video itself. Based on CNNs, the proposed method is devised to classify videos based on visual and aural information that can be extracted from the content of the video. The visual content of a video is derived from patches cropped from its video frames and the audio content is derived from patches cropped from the audio track's Log-Mel Spectrogram. To classify the query video, we use the Late Fusion method where we combine the scores obtained from two mono-modal networks (one working with visual patches and the other working with audio patches), and feed them into one multi-input network with visual/audio patch pairs extracted from the query video. The Early Fusion method uses a single multi-input network that is fed by visual/audio patch pairs extracted from the query video. It is important to note that both of these approaches are multi-modal methods of identifying camera models. Our study aims to examine three different topologies for each approach, with the use of various architectures and data pre-processing methods to do so. Using video clips that were taken from the Vision dataset, we assess the effectiveness of our experimental campaign. The videos we test are not just the original native ones that were captured by the smartphone camera directly, but we also test other videos as well. The purpose of this videos is to investigate a variety of training and testing configurations, as well as to come up with a way to simulate real-world scenarios in which it is necessary for us to categorize data compressed through internet services. In order to achieve these goals, we also use movies compressed using WhatsApp and YouTube algorithms (for example, social media, and upload sites). In addition, we compare the multi-modal attribution strategy we propose to the traditional mono-modal attribution strategy as well as other suggested techniques [22]. On average, the Late Fusion technique provides the best outcomes of the var- ious multi-modal approaches and significantly outperforms traditional mono-modal approaches; the data confirms that the multi-modal approaches outperform mono- modal approaches. There are generally fewer than 99 percent chances that we will be able to correctly distinguish an original video sequence from a YouTube video sequence.99 percent. There are still some videos that are difficult to model, mainly because of the extreme compression used in WhatsApp, which may have something to do with the difficulty. It is obvious that this opens up possibilities for new problems and advancements centered around the identification of the originating camera model for videos that are posted (or shared repeatedly) on social media. Additionally, it is important to note that the suggested multi-modal solutions can be applied easily to a hypothetical situation where there are more than two data modalities being used. As a result of using the Late Fusion approach, the CNNs would only have to be trained independently on each target”. When films share sequential data, one potential option would be to look into how neighbouring frames might be utilized for scene suppression and boosting the separation of camera noise [10].

References

[1] Abdali, S. (2022). Multi-modal Misinformation Detection: Approaches, Challenges and Opportunities. http://arxiv.org/abs/2203.13883

[2] Abdullakutty, F., Johnston, P., & Elyan, E. (2022). Fusion methods for Face Presentation Attack Detection. https://doi.org/10.3390/s22145196

[3] Akbari, Y., Al-Maadeed, S., Al-Maadeed, N., Najeeb, A. A., Al-Ali, A., Khelifi, F., & Lawgaly, A. (2022). A New Forensic Video Database for Source Smartphone Identification: Description and Analysis. IEEE Access, 10, 20080–20091. https://doi.org/10.1109/ACCESS.2022.3151406

[4] Akilan, T., Jonathan Wu, Q. M., Jiang, W., Safaei, A., & Huo, J. (2019). New trend in video foreground detection using deep learning. Midwest Symposium on Circuits and Systems, 2018-August (Cv), 889–892. https://doi.org/10.1109/MWSCAS.2018.8623825

[5] Al Banna, M. H., Ali Haider, M., Al Nahian, M. J., Islam, M. M., Taher, K. A., & Kaiser, M. S. (2019). Camera model identification using deep CNN and transfer learning approach. 1st International Conference on Robotics, Electrical and Signal Processing Techniques, ICREST 2019, January, 626–630. https://doi.org/10.1109/ICREST.2019.8644194

[6] Amerini, I., Anagnostopoulos, A., Maiano, L., & Celsi, L. R. (2021). Deep Learning for Multimedia Forensics. In Deep Learning for Multimedia Forensics. https://doi.org/10.1561/9781680838558

[7] Ashraf, A., Gunawan, T. S., Riza, B. S., Haryanto, E. V., & Janin, Z. (2020). On the review of image and video-based depression detection using machine learning. Indonesian Journal of Electrical Engineering and Computer Science, 19(3), 1677–1684. https://doi.org/10.11591/ijeecs.v19.i3.pp1677-1684

[8] Athanasiadou, E., Geradts, Z., & Van Eijk, E. (2018). Camera recognition with deep learning. Forensic Sciences Research, 3(3), 210–218. https://doi.org/10.1080/20961790.2018.1485198

[9] Bennabhaktula, G. S., Timmerman, D., Alegre, E., & Azzopardi, G. (2022). Source Camera Device Identification from Videos. SN Computer Science, 3(4), 1–15. https://doi.org/10.1007/s42979-022-01202-0.

[10] Bhatti, M. T., Khan, M. G., Aslam, M., & Fiaz, M. J. (2021). Weapon Detection in Real-Time CCTV Videos Using Deep Learning. IEEE Access, 9, 34366–34382. https://doi.org/10.1109/ACCESS.2021.3059170

[11] Blasch, E., Liu, Z., & Zheng, Y. (2022). Advances in deep learning for infrared image processing and exploitation. May, 56. https://doi.org/10.1117/12.2619140

[12] Ciaparrone, G., Luque Sánchez, F., Tabik, S., Troiano, L., Tagliaferri, R., & Herrera, F. (2020). Deep learning in video multi-object tracking: A survey. Neurocomputing, 381, 61–88. https://doi.org/10.1016/j.neucom.2019.11.023

[13] Dal Cortivo, D., Mandelli, S., Bestagini, P., & Tubaro, S. (2021). CNN-based multi-modal camera model identification on video sequences. Journal of Imaging, 7(8). https://doi.org/10.3390/jimaging7080135

[14] Fan, H., Murrell, T., Wang, H., Alwala, K. V., Li, Y., Li, Y., Xiong, B., Ravi, N., Li, M., Yang, H., Malik, J., Girshick, R., Feiszli, M., Adcock, A., Lo, W. Y., & Feichtenhofer, C. (2021). PyTorchVideo: A Deep Learning Library for Video Understanding. MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia, 3783–3786. https://doi.org/10.1145/3474085.3478329

[15] Gona, A., & Subramoniam, M. (2022). Convolutional neural network with improved feature ranking for robust multi-modal biometric system. Computers and Electrical Engineering, 101(November 2021), 108096. https://doi.org/10.1016/j.compeleceng.2022.108096

[16] Guera, D., Wang, Y., Bondi, L., Bestagini, P., Tubaro, S., & Delp, E. J. (2017). A Counter-Forensic Method for CNN-Based Camera Model Identification. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2017-July, 1840–1847. https://doi.org/10.1109/CVPRW.2017.230

[17] Hosler, B., Mayer, O., Bayar, B., Zhao, X., Chen, C., Shackleford, J. A., & Stamm, M. C. (2019). A Video Camera Model Identification System Using Deep

Learning and Fusion. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2019-May, 8271–8275. https://doi.org/10.1109/ICASSP.2019.8682608

[18] Huynh, V. N., & Nguyen, H. H. (2021). Fast pornographic video detection using Deep Learning. Proceedings - 2021 RIVF International Conference on Computing and Communication Technologies, RIVF 2021. https://doi.org/10.1109/RIVF51545.2021.9642154

[19] Jagannath Patro, S., & M, N. V. (2019). Real Time Video Analytics for Object Detection and Face Identification using Deep Learning. 8(05), 462–467. www.ijert.org

[20] Maiano, L., Amerini, I., Ricciardi Celsi, L., & Anagnostopoulos, A. (2021). Identification of social-media platform of videos through the use of shared features. Journal of Imaging, 7(8). https://doi.org/10.3390/jimaging7080140

[21] Member, S., & Member, S. (2021). MMHAR-EnsemNet : A Multi-Modal Human. 21(10), 11569–11576.

[22] Ott, J., Atchison, A., Harnack, P., Bergh, A., & Linstead, E. (2018). A deep learning approach to identifying source code in images and video. Proceedings - International Conference on Software Engineering, 376–386. https://doi.org/10.1145/3196398.3196402

[23] Pandeya, Y. R., & Lee, J. (2021). Deep learning-based late fusion of multimodal information for emotion classification of music video. Multimedia Tools and Applications, 80(2), 2887–2905. https://doi.org/10.1007/s11042-020-08836-3

[24] Phan, T., Phan, A., & Cao, H. (2022). applied sciences Content-Based Video Big Data Retrieval with Extensive Features and Deep Learning. 1–26.

[25] Ramos Lopez, R., Almaraz Luengo, E., Sandoval Orozco, A. L., & Villalba, L. J. G. (2020). Digital video source identification based on container's structure analysis. IEEE Access, 8, 36363–36375. https://doi.org/10.1109/ACCESS.2020.2971785

[26] Salido, J., Lomas, V., Ruiz-Santaquiteria, J., & Deniz, O. (2021). Automatic handgun detection with deep learning in video surveillance images. Applied Sciences (Switzerland), 11(13). https://doi.org/10.3390/app11136085

[27] Schofield, D., Nagrani, A., Zisserman, A., Hayashi, M., Matsuzawa, T., Biro, D., & Carvalho, S. (2019). Chimpanzee face recognition from videos in the wild using deep learning. Science Advances, 5(9), 1–10. https://doi.org/10.1126/sciadv.aaw0736

[28] Shi, Y., & Biswas, S. (2019). A Deep-Learning Enabled Traffic Analysis Engine for Video Source Identification. 2019 11th International Conference on Communication Systems and Networks, COMSNETS 2019, 2061, 15–21. https://doi.org/10.1109/COMSNETS.2019.8711478

[29] Shi, Y., Feng, D., Cheng, Y., & Biswas, S. (2021). A natural language-inspired multilabel video streaming source identification method based on deep neural networks. Signal, Image and Video Processing, 15(6), 1161–1168. https://doi.org/10.1007/s11760-020-01844-8

[30] Shojaei-Hashemi, A., Nasiopoulos, P., Little, J. J., & Pourazad, M. T. (2018). Video-based Human Fall Detection in Smart Homes Using Deep Learning. Proceedings - IEEE International Symposium on Circuits and Systems, 2018-May, 0–4. https://doi.org/10.1109/ISCAS.2018.8351648

[31] Sreenu, G., & Saleem Durai, M. A. (2019). Intelligent video surveillance: a review through deep learning techniques for crowd analysis. Journal of Big Data, 6(1), 1–27. https://doi.org/10.1186/s40537-019-0212-5

[32] Uddin, M. A., Joolee, J. B., & Sohn, K. A. (2022). Deep Multi-Modal Network Based Automated Depression Severity Estimation. IEEE Transactions on Affective Computing, 14(8). https://doi.org/10.1109/TAFFC.2022.3179478

[33] Wang, W., Li, X., Xu, Z., Yu, W., Zhao, J., Ding, D., & Chen, Y. (2022). Learning Two-Stream CNN for Multi-Modal Age-related Macular Degeneration Categorization. IEEE Journal of Biomedical and Health Informatics, X(X), 1–12. https://doi.org/10.1109/JBHI.2022.3171523

[34] Wang, Y., Sun, Q., Rong, D., Li, S., & Xu, L. Da. (2021). Image Source Identification Using Convolutional Neural Networks in IoT Environment. Wireless Communications and Mobile Computing, 2021. https://doi.org/10.1155/2021/5804665

[35] Wodajo, D., & Atnafu, S. (2021). Deepfake Video Detection Using Convolutional Vision Transformer. http://arxiv.org/abs/2102.11126

[36] Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X. (2019). Object Detection with Deep Learning: A Review. IEEE Transactions on Neural Networks

AHP Algorithm for Indoor Air Pollution Detection and Evaluation System Design

Wen Fan[1*], Chengyang Chang[2], Ning Yao[1], Linxue Xu[1], Hongyan Ju[1]
[1]Cangzhou Jiaotong College, HuangHua, HeBei, 061199, China
[2]Tianjin University of Technology and Education, TianJin, 300350, China
E-mail: wenfan291@163.com, chengyangchang7@126.com, ningyao71@163.com, linxuexu8@163.com, hongyanju9@163.com

Keywords: air quality, closed environment, pollution detection, evaluation system design, AHP algorithm

Received: January 24, 2022

Abstract: The superiority of buildings, considering their routine in a variety of indoor environmental qualities, is significant to the existing habitants potentials. In order to extract a description of indoor air quality, the concentration of indoor pollutants needs to be obtained and then evaluated. Aiming at the existing indoor air quality monitoring and evaluation system, an AHP algorithm for the design of indoor air pollution detection and evaluation system is proposed, which combines the principles and methods of fuzzy mathematics to evaluate air quality in a confined environment. An experiment was carried out, using the analytic hierarchy process to assign weights. According to the principle of maximum subordination, comprehensive evaluation of air quality in a confined environment was carried out through fuzzy mathematical model. The results from experiments demonstrates the scope of improvement in the design of new buildings and work with prioritization of restricted assets for updating the performance of building. The experimental results show that the humidity value reached 54% RH when the temperature was 25°C, and the humidity value reached 60% RH when the temperature was 19°C. The evaluation results more scientifically reflect the true state of air quality.

Povzetek: Predlagan je sistem za oceno kakovosti zraka v zaprtih prostorih z uporabo AHP algoritma in metode mehke matematike.

1 Introduction

With the improvement of people's living standards, people have paid great attention to environmental pollution [1]. Environmental pollution has also become more important to people. The harm caused by the environmental pollution of the British Industrial Revolution and the environmental pollution problems that the United States once faced have brought heavy lessons for today’s fast-developing science and technology in China. Faced with China's current state of affairs, the development of science and technology is bound to cause damage to the environment. The destruction of outdoor air quality will definitely affect the indoor environment of people's lives. According to a World Bank statistical report, China’s annual economic loss of approximately US $3.2 billion is due to indoor environmental pollution. In addition, according to the investigation of the international testing and environmental agency, at least 30% of the indoor environments of buildings in the world contain harmful substances that endanger human health. The quality of indoor air directly affects people's health and living standards. Indoor environmental pollution has been listed as one of the five environmental factors that are the most harmful to public health [2].

Skyscraper high rises are packed all over China due to congestion. As of late, the load of public and confidential lodging expanded by 782,754 and 1,634,847, separately [3]. Minimal expense public lodging pads have average designs and are typically little, ready to oblige private examples and inclinations. Confidential pads are normally involved by center pay proprietors or occupants. Albeit the living quarters of these pads are bigger, the normal cruciform construction joined with a focal plan lessens the nature of public spaces in those structures [4]. Aside from the level size and design, numerous different factors, for example, the outer climate [5], closeness to foundation and assets and the elements of indoor lodging, for example, family room influence the living space of occupants. The multi-faceted natural execution of the structure, regarding the nature of the indoor environment, perceivability and sound and the nature of the indoor air, additionally affect the strength of inhabitants and their fulfillment with the living space [6].

To work on the presentation of structures in the climate, various deliberate investigation programs have arisen, for example, the Building Research Establishment's Environmental Assessment Method (BREEAM) in the UK and in China, the HK-BEAM Organization has had huge progress in expanding enlistment in its Voluntary China Building Environmental Assessment Program [7]. These cycles frequently incorporate the assessment of the quantity of indoor climate (IEQ); each with a particular credit point for the general outcome. Progressively, the logical order (AHP) process is utilized to handle the speculations, to decide the upsides of the connection between's the tried factors, which are expected to consolidate the joined test results into accumulated places (for example [8, 9]). Concentrates on directed on business structures found that different psychophysical factors impact basic judgment on saw esteem [10]. In private structures, be that as it may, little is had some significant awareness of the occupants' impression of the significance of legitimacy [11]. While property executive organizations might lead a general review of inhabitant’s fulfilment with their living space, for example the apparent usefulness of structures under their administration, to recognize regions for improvement, there is an absence of rules on the most proficient method to focus on so more prominent improvement can be accomplished reasonably affordable. The review was expected to close this data hole in the China context. The manuscript is composed as follows: Section 2 is for related work and Section 3 is for research methodology. Section 3 includes research methods followed by results and analysis in Section 4. Last section includes conclusion.

2 Related work

In this section various state-of-the-art work in the field of indoor air pollution detection and evaluation presented.

With the development of society, Nag *et al.* found that modern people spend about 80% of their lives indoors, and people's life, work, entertainment and other activities are concentrated indoors [12]. According to survey data, Zacarías *et al.* found that there are 7 million direct or indirect deaths caused by indoor air pollution each year, of which China accounts for one-seventh of the total deaths. Indoor air pollution can be divided into chemical pollution, physical pollution, biological pollution and radioactive pollution. Chemical pollution mainly includes volatile compounds and inorganic gases such as formaldehyde (CH0) [13], carbon monoxide (C0) and carbon dioxide (CO2). Lee *et al.* found that physical pollution mainly includes inhalable particles, dust, etc.; biological pollution is mainly caused by biological fungi, bacteria and other microorganisms; radioactive pollution is mainly radioactive substances remaining in the indoor air [14]. Xie *et al.* found that the four types of pollutants such as oxygen Ra affect human sensory experience and even physical health, and have a great impact on the human respiratory system, endocrine system, and nervous system [15], and even cause disease.

The main feature of indoor pollutants is that they are exposed to a wide range of people, and different people have different sensitivity, age and health factors. In addition, there are many pollutants, such as the daily inhalable particulate matter PM2.5 and PM10, which can directly enter the human lungs and cause extensive lung fibers, leading to pneumoconiosis. CH20 is highly toxic and volatile, which is extremely harmful to the human body. Miao *et al.* found that CO and CO2 are inorganic gaseous pollutants. Normal CO2 concentration has no obvious effect on the human body. However, if the concentration is too high, people will have symptoms such as sleepiness and lack of energy. CO is insufficient carbon. The product formed by combining with oxygen [16] is colorless, odorless, and highly toxic. Deeply poisoned, it will cause irreversible and permanent damage to the brain. Long-term exposure to these pollutants can cause harm to the human body and cause disease. Oyabu *et al.* understand the concentration of various indoor pollutants and judge the pollution level of the indoor environment [17].

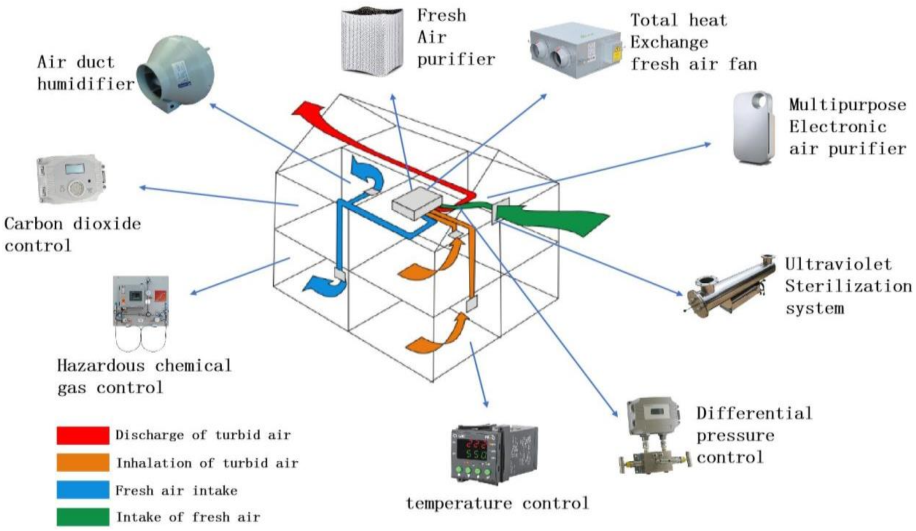

Air duct humidifier
Fresh Air purifier
Total heat Exchange fresh air fan
Multipurpose Electronic air purifier
Carbon dioxide control
Ultraviolet Sterilization system
Hazardous chemical gas control
Differential pressure control
temperature control
🟥 Discharge of turbid air
🟧 Inhalation of turbid air
🟦 Fresh air intake
🟩 Intake of fresh air

Figure 1: Indoor air quality system solution.

It can keep away from the heavily polluted area in time to avoid harm to the human body. It is very important to monitor these pollutants in real time and evaluate the pollution level. With the rapid development of Internet of Things technology in recent years, China has established a system to monitor outdoor air.

The Environmental Protection Agency will also publish the air environmental quality index and pollution level of each region. Oyabu *et al.* found that it is possible to query the specific concentration values and pollution levels of outdoor air pollutants in real time. At present, most of the indoor air quality monitoring systems are based on wireless sensor networks, such as AHP. WIFI, etc. 18]. Wu *et al.* found that the wireless sensor network itself has its own limitations, requiring the deployment of a wireless office network, which brings a lot of inconvenience. When the monitoring points are many and the distribution range is wide, such as monitoring the air quality status of all residents in a certain community, or even monitoring multiple communities, wireless sensor network deployment is difficult, and there are higher requirements for the deployment network connection points. The traditional sensor network is difficult to implement, and can't even meet the demand [19]. Bluyssen and Cox found that based on this choice, it is also very important to be able to connect a large number of wireless communication technologies with a large coverage area, and a more intelligent and convenient monitoring system is needed to meet the needs of different indoor monitoring scenarios [20].

Mirmohammadi *et al.* found that currently China has not issued a clear and relevant grade evaluation standard, which has caused different departments to use different evaluation standards, and the evaluation results obtained are also different [21]. According to the information obtained, most of the current indoor air quality evaluation systems mainly focus on single-factor factors, such as the evaluation of indoor inhalable particulate pollution and the evaluation of formaldehyde pollution. Some parts have been comprehensively evaluated, but the pollution factors and the evaluation methods used are not the same. In Figure, it has been dictated that how to choose appropriate pollution influencing factors and evaluation methods to obtain a reasonable evaluation level, and comprehensively and objectively reflect the current monitoring indoor air quality.

3 Research methods

This section describes the adopted methodology for the detection and evaluation of indoor air pollution. The air quality evaluation of a closed environment is to analyze, evaluate and predict the air quality of a certain enclosed area according to certain evaluation standards and evaluation methods [22]. At present, there are many air quality evaluation methods, and the comprehensive pollution index method and expert scoring evaluation method are commonly used. However, in the evaluation process of various situations, due to the many evaluation factors and no clear indicators, the evaluation has a certain tendency and is not objective, and the comprehensive evaluation results have a certain deviation. Especially in the actual confined environment, the impact of human activities and equipment operation on the air quality of the confined environment is more complicated, and the comprehensive effect is difficult to determine. In order to obtain reasonable air environmental quality evaluation results, more and more fuzzy theories have been introduced in recent years to deal with this transitional gradual problem. The theory uses AnaVtic HienarthyProcess AHP to determine the weight of each factor, establishes a fuzzy comprehensive evaluation model, and obtains a more objective evaluation result [23].

It uses the 1~9 ratio scale method suggested by SAATY to construct a pairwise comparison judgment matrix. If there exists $a_{ij} = a_{ik}$ relationship, the matrix is said to have complete consistency. The eigenvector corresponding to its largest eigenvalue can give the relative importance of the index. The order, after orthogonalizing it, is the desired weight vector.

By calculating the maximum eigenvalue λ_{max} consistency index CI and consistency ratio CR, to check whether the consistency of the comparison matrix established above meets the requirements. After determining the fuzzy matrix R and the weight A of each factor, the fuzzy comprehensive evaluation model of the overall environmental quality can be obtained through its compound operation U=A°R. There are many kinds of fuzzy calculation methods, here we use “-” and “+” operators, denoted as M(.,+). In this model, since each factor is normalized, the operation + degenerates into a general real number addition. And this model considers the influence of all factors when determining the degree of membership of the evaluation of each factor to the grade, and the calculation is more refined. Then the elements in U are presented in Equation 1.

$$U_{ij} = \min\left(1, \sum_{1}^{n} a_{ik} . \quad b_{kj} = \min\{1, [a_1 \cdot b_{ij}] + (a_2 \cdot b_2) + \ldots (a_k \cdot b_k)\right) \qquad (1)$$

Table 1: Sub-index of mass concentration of sampling items

Sampling point value	C6H6C6H6		HOCO	O $O_2O_2H_3O_2$
1#	5.72	33.5	12.6	.11.03.62.13.61
2#	2.51	12.4	0.87	.42.88.86.53.41
3#	1.97	22.4	4.12	.56.96.62.08.61
Average value	3.41	22.86	5.82	.26.63.02.59.3

When evaluating air quality in a closed environment, pollutants with different levels of harm to the human body are selected as the evaluation objects, and the original data of sampling points are selected. After statistical sorting, the relative value is used for fuzzy processing (the ratio of the average value of the measured concentration to the allowable concentration), see Table 1.

With reference to the allowable concentration and emergency allowable concentration of the closed environment, the air quality of the closed environment allows the people working and living within the allowable specified time, and the air quality grade index of the closed environment is defined as: the ratio of each emergency allowable concentration of air quality evaluation parameters in a closed environment to the allowable concentration of the closed environment. The air quality level of the closed environment is divided into 4 levels, which respectively represent the four levels of clean, light pollution, medium pollution and heavy pollution in the air quality status, that is, the evaluation set $V = \{I, II, I, IV\}$. The sub-indices of the air quality classification standards for airtight environments are shown in Table 2.

Table 2: Sub-indexes of air quality classification standards for confined environments

Level	**C6H6**	**Hg**	**HOCO**	**CO**	**SO_2**	**NO_2**	**H_3**	**O_2**
Level I	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01
Level II	60.2	43.85	13.32	8.6	13.01	6.66	3.56	3.01
Level III	13.20	51.13	26.66	21.73	26.02	20.01	5.11	4.01
Level IV	20.03	142.85	40.01	39.01	40.07	41.30	8.01	4.40

Table 3: Single factor evaluation results

Hierarchical membership	**C6H6**	**Hg**	**HOCO**	**CO**	**SO_2**	**NO_2**	**NH_3**	**CO_2**
Level I (clean)	1.02	1.02	1.02	1.02	1.02	1.02	1.02	1.02
Level II (light pollution)	60.1	42.85	12.32	8.5	13.05	6.65	3.55	3.02
Level III(medium pollution)	13.1	51.12	26.65	21.72	26.01	20.03	5.16	4.05
Level IV(heavy pollution)	20.02	142.86	41.01	39.01	40.06	41.2	8.02	4.30

Figure 2 presents the computation of important weights and consistency ratios. Initially, from input each set of data is organized in 4 × 4 comparison matrix. Then in next step the organized data is forwarded to the standard program for the calculation of Eigen values and Eigen vectors. In next step the principal Eigen values are extracted and therefore consistency ratio is computed. In next step the principal Eigen vectors are normalized and then the consistency ratio is evaluated. After the evaluation, normalized Eigen vectors are accepted and the inconsistent dataset is discarded.

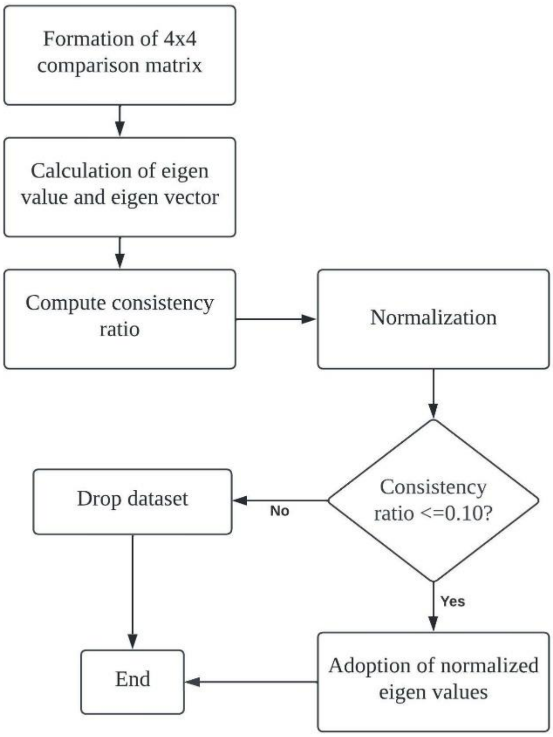

Formation of 4x4 comparison matrix
Calculation of eigen value and eigen vector
Compute consistency ratio
Normalization
Consistency ratio <=0.10?
No
Drop dataset
Yes
Adoption of normalized eigen values
End

Figure 2: Computation of weight and consistency ratio.

It can be seen that the comprehensive assessment result of the air quality in the enclosed environment is that the membership degree of the air environment quality to the I, II, II, and IV levels is 0.4271.5310.0420, respectively. The single factor evaluation results are presented in Table 3. According to the principle of maximum affiliation, it is comprehensively assessed as a secondary standard (light pollution). The fuzzy mathematics method is used to establish an evaluation model for air quality evaluation in a closed environment, and the weight coefficient is determined by the analytic hierarchy process (AHP), and the comprehensive evaluation result of the air quality in this closed environment is obtained as the level II standard (light pollution). Using fuzzy mathematics to describe the air quality status of a closed environment with a membership function can not only get the air environment quality level, but also reflect the membership status of various pollutants, which improves the scientific nature of the evaluation. However, if the membership function is not established properly, the unreasonable setting of the weight function will also cause inaccurate evaluation results. Therefore, it needs to be further improved in practical application to establish a more scientific and reasonable model and evaluation method.

4 Results and analysis

This section illustrates the analysis of results obtained by comparing the seismic forces and finally presents its discussion and summary.

In order to verify the accuracy of the formaldehyde detection and the performance of the formaldehyde detection module, the formaldehyde module is connected to the corresponding pins of the main control chip, and the formaldehyde concentration detected by the detection system is displayed and recorded in real time. The working principle of the formaldehyde sensor is a two-electrode electrochemical sensor, which realizes the detection of formaldehyde through the principle of diffusion. Since the sensor used in this detection system is a mature electrochemical sensor developed in China, the Chengsan company has made a specific calibration statement for the corresponding calibration of its formaldehyde concentration before it is put into use. The sensor can be put into use directly. When detecting and calibrating the formaldehyde gas, the formaldehyde gas detected by the sensor is mainly detected from the perspective of the sensitivity, stability, response time, calibration curve, and experimental data of the formaldehyde sensor.

A. Sensitivity

The sensitivity of a formaldehyde sensor refers to the lowest value that a certain sensor can detect the concentration of formaldehyde. The factors that affect the sensitivity are affected by the diffusion rate of formaldehyde gas and electrolyte in the sensor and the chemical characteristics of the internal working electrode [38]. The sensitivity of the ME3-CH2O sensor is defined as: sensitivity=2000nA/cm3.

B. Stability

When measuring the stability of the sensor, two values need to be measured: zero drift and span drift. The following table 4 shows the regulations for the stability of the ME3-CH2O sensor:

Zero drift ≤ 10% FS Span drift ≤ 10% FS

C. Response time

Response time is a performance indicator that reflects the speed of the sensor. The sensing speed of the electrochemical formaldehyde sensor is determined by the electrolyte resistance between the working electrode and the reference electrode. The sensor used in this system has a well-defined response time in the design, and its stipulation is: response time ≤ 30S.

D. Calibration curve of the senor

The formaldehyde sensor uses the method of permeation tube to define the calibration curve of the sensor. The theoretical basis is that the volatile isolation layer between the water and the top air is constant at a specially designated temperature. The calibration curve is provided for testing in 1.0ppm formaldehyde at 20°C. The following calibration curve of 1.0ppm formaldehyde solution at 20°C is shown in Figure 3.

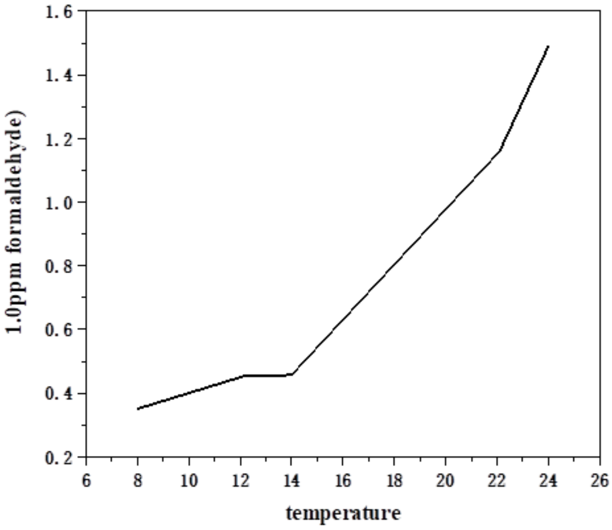


Figure 3: Calibration curve of HOCO.

E. Experimental data

In order to detect whether the measurement of formaldehyde gas is accurate, it was calibrated by comparing it with other formaldehyde gas sensors. The current application WP6900 Agris professional formaldehyde sensor was selected for calibration. According to the relevant information of WP6900, the minimum resolution of the detection device is 0.001mg/cm’. The minimum resolution designed by this detection system is 0.01mg/cm'. Because the resolution of this detection design is lower than that of the WP6900 detection device, it can be calibrated.

Table 4: System test results

Location	**Temperature**	**Humidity**	**HOCO concentration**	**State**
Laboratory	23℃	49％RH	0.03mg/m^3	Normal
Bedroom	25℃	51％RH	0.02mg/m^3	Normal
Canteen	17℃	64％RH	0.03mg/m^3	Normal

A certain concentration of gaseous formaldehyde was obtained by heating the liquid formaldehyde solution for gas detection. The formaldehyde generator can obtain different concentrations of gaseous formaldehyde by heating a certain concentration of liquid formaldehyde and calculating through the formula of the experimental device. The corresponding concentration of gaseous formaldehyde was obtained by heating different concentrations of liquid formaldehyde for measurement. The experiment selected 5 kinds of gaseous formaldehyde with different concentrations, and the calculated concentration of formaldehyde was 0.01mg/cm', 0.015mg/cm3, 0.02mg/cm3, 0.025mg/cm', 0.03mg/cm3. The measured environment was an indoor temperature of 25°C. The HOCO concentration values detected by the two sensors were recorded, as shown in Table 5.

Table 5: Laboratory testing data of HOCO

Production of experimental device	**The detection system**	**WP6900**
0.01mg/cm^3	0.01mg/cm^3	0.008mg/cm^3
0.014mg/cm^3	0.03mg/cm^3	0.014mg/cm^3
0.03mg/cm^3	0.03mg/cm^3	0.020mg/cm^3
0.024mg/cm^3	0.04mg/cm^3	0.023mg/cm^3
0.02mg/cm^3	0.04mg/cm^3	0.026mg/cm^3

Due to the simple laboratory equipment and the standard gas environment is not easy to reach, formaldehyde gas is dangerous to a certain extent. Therefore, there is a certain error when testing and calibrating in the laboratory. The values measured by the two sensors are compared; the results are shown in Figures 4 and 5.

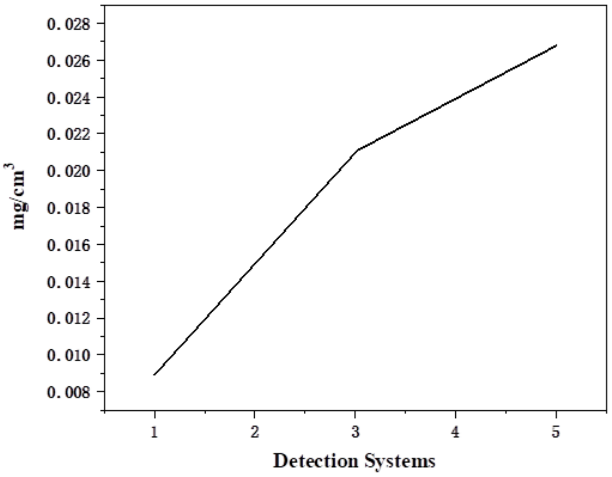


Figure 4: Formaldehyde concentration detected by WP6900.

By comparing the flow chart of the detection results, it can be seen that the designed detection device and the WP6900 have a difference in resolution due to the difference between the two curves, but the detection system can basically meet the measurement of formaldehyde gas concentration.

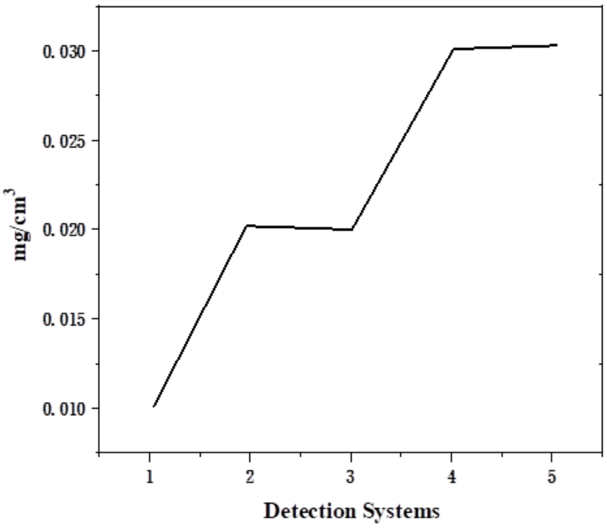


Figure 5: Formaldehyde concentration detected by the detection system.

The performance is measured in terms of indoor environmental quality for living and common area through local residents and visitors. Figure 6 represents the performance measurement of indoor environmental quality rated by residents. Figure 6 represents the performance measurement of indoor environmental quality rated by visitors. The visitors are ranked higher in comparison with the residents for the performance measurement. For example, all the four found the middle value of execution appraisals for the four indoor environmental quality credits given by the guests for the living/visited region in confidential structures surpass 5.1 yet the most noteworthy arrived at the midpoint of rating given by the occupants is lower than 5.0 as presented in Figure 6 and 7. Comparative connection exists for living/visited region in open structures, however not for all indoor environmental quality attributes.

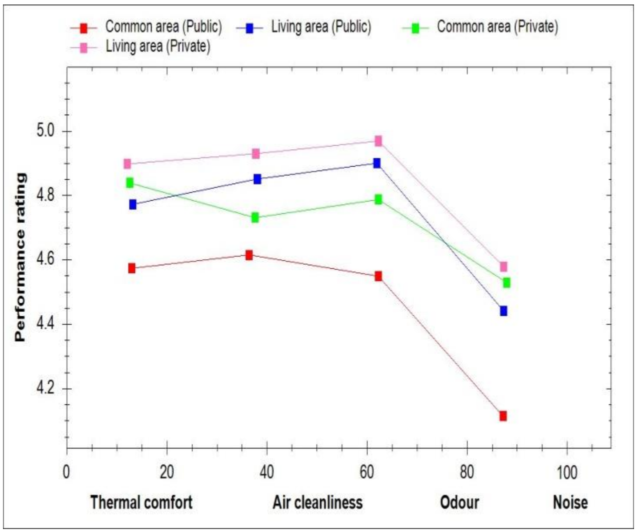


Figure 6: Performance measurement of indoor environmental quality rated by residents.

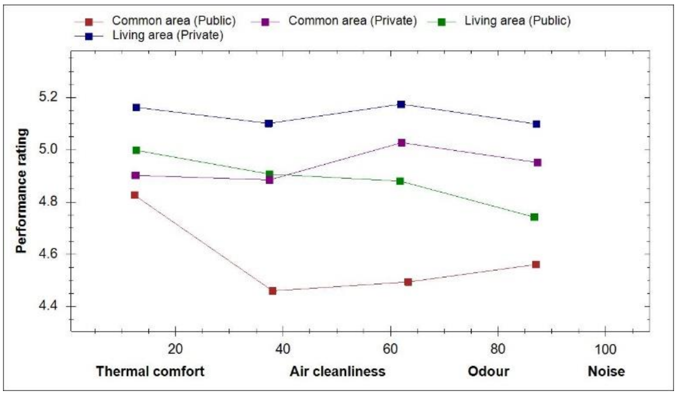


Figure 2: Performance measurement of indoor environmental quality rated by visitors.

The temperature sensor DS18B20 and HS1101 humidity detection module to the processor was connected through the corresponding interface. The detected data is processed and sent, and printed and displayed on the LCD screen and serial port, so that the temperature and humidity of the greenhouse can be measured. The main control chip can complete the measurement and display of temperature and humidity by using the microcontroller to code the temperature detection and humidity detection modules. The detected humidity value was displayed on the LCD screen or printed to the serial port. The purpose was to allow us to see the detected temperature and humidity value more intuitively. The indoor greenhouse temperature values detected by the detection module are shown in Table 6.

Table 6: Formaldehyde concentration detected by the detection system

Temperature value	Humidity value
25℃	50％RH
23℃	49％RH
17℃	64％RH
19℃	60％RH

In the software programming of temperature and humidity detection, accurate and stable temperature and humidity values are obtained, and finally output is displayed in order to provide convenient applications for people.

5 Conclusions

A low-power, portable indoor air quality detection system based on AHP algorithm IS designed. In the process of designing and completing the whole detection system, the research status of indoor air quality detection system was analyzed in detail, and the overall structure of the system was designed. According to the hardware platform built by the detection system, the detection system software scheme is designed to make the hardware part of the detection system complete the system requirements. The humidity value reaches 54%RH at 25℃, and the humidity value reaches 60%RH at 19℃. e the results of the designed detection system is tested and analyzed. It is mainly for testing the detection module, the communication test between the system detection device and the intelligent terminal upper computer, the overall function of the system, and the analysis of the test results. Although the designed indoor air quality detection system can achieve its basic detection functions, the system still has shortcomings and improvements. Due to time and condition constraints, this design still needs further improvement. Since the designed indoor air detection system is still in the laboratory stage, the mutual application of the functional modules used by the detection system will increase the cost and power consumption of the detection system. The purpose of reducing system power consumption and cost can be achieved by designing a reasonable application circuit and reducing the size of the board. The performance analysis in the normal region that the clients perceived was for the most part lower than the corresponding in the living/visited region, among which the worst thing is noticed as noise. The performance examination, in light of the general reaction of the clients, has shown the way that the gaps between them can be distinguished. This is the sort of data that helps managers for improving the utilization of the frequently compelled assets to manage structures. Besides, it can illuminate the regions for development in existing structures and the vital adjustments for building plan in future.

Acknowledgement:

Project name of Cangzhou Science and Technology Bureau: Design and Implementation of Indoor Environmental Detection System Based on the Internet of Things (No.: 183103006)

References

[1] Park, J. Y., Kim, N., & Shin, Y. K. (2016). Design of Pitch Limit Detection Algorithm for Submarine. *Journal of Ocean Engineering and Technology*, *30*(2), 134-140. *https://doi.org/10.5574/KSOE.2016.30.2.134*

[2] Manickam, C., Raman, G. P., Raman, G. R., Ganesan, S. I., & Chilakapati, N. (2016). Fireworks enriched P&O algorithm for GMPPT and detection of partial shading in PV systems. *IEEE Transactions on Power Electronics*, *32*(6), 4432-4443. *10.1109/TPEL.2016.2604279*

[3] Yu, W., Li, B., Yang, X., & Wang, Q. (2015). A development of a rating method and weighting system for green store buildings in China. *Renewable Energy*, *73*, 123-129. https://doi.org/10.1016/j.renene.2014.06.013

[4] Cho, S. H., Lee, T. K., & Kim, J. T. (2011). Residents' satisfaction of indoor environmental quality in their old apartment homes. *Indoor and Built Environment*, *20*(1), 16-25. https://doi.org/10.1177/1420326X10392010

[5] Chandratilake, S. R., & Dias, W. P. S. (2013). Sustainability rating systems for buildings: Comparisons and correlations. *Energy*, *59*, 22-28. https://doi.org/10.1016/j.energy.2013.07.026

[6] Zheng, Q., Lee, D., Lee, S., Kim, J. T., & Kim, S. (2011). A health performance evaluation model of apartment building indoor air quality. *Indoor and Built Environment*, *20*(1), 26-35. https://doi.org/10.1177/1420326X10393719

[7] Karaca, F. (2015). An AHP-based indoor Air Pollution Risk Index Method for cultural heritage collections. *Journal of cultural Heritage*, *16*(3), 352-360. https://doi.org/10.1016/j.culher.2014.06.012

[8] Ali, H. H., & Al Nsairat, S. F. (2009). Developing a green building assessment tool for developing countries–Case of Jordan. *Building and environment*, *44*(5), 1053-1064. https://doi.org/10.1016/j.buildenv.2008.07.015

[9] Yuan, J., Chen, Z., Zhong, L., & Wang, B. (2019). Indoor air quality management based on fuzzy risk assessment and its case study. *Sustainable Cities and Society*, *50*, 101654. https://doi.org/10.1016/j.scs.2019.101654

[10] Wong, J. K., & Li, H. (2008). Application of the analytic hierarchy process (AHP) in multi-criteria analysis of the selection of intelligent building systems. *Building and Environment*, *43*(1), 108-125. https://doi.org/10.1016/j.buildenv.2006.11.019

[11] Heinzerling, D., Schiavon, S., Webster, T., & Arens, E. (2013). Indoor environmental quality assessment models: A literature review and a proposed weighting and classification scheme. *Building and environment*, *70*, 210-222. https://doi.org/10.1016/j.buildenv.2013.08.027

[12] Nag, D., Paul, S. K., Saha, S., & Goswami, A. K. (2018). Sustainability assessment for the transportation environment of Darjeeling, India. *Journal of environmental management*, *213*, 489-502. *https://doi.org/10.1016/j.jenvman.2018.01.042*

[13] Zacarías, S. M., Manassero, A., Pirola, S., Alfano, O. M., & Satuf, M. L. (2021). Design and performance evaluation of a photocatalytic reactor for indoor air disinfection. *Environmental Science and Pollution Research*, *28*(19), 23859-23867. https://doi.org/10.1007/s11356-020-11663-6

[14] Lee, J. Y., Chung, Y. S., Kim, D. S., Bae, G. H., Bae, J. S., & Lee, D. H. (2017). Development of weft straightener using fabric pattern detection algorithm and performance evaluation. *Korean*

Journal of Computational Design and Engineering, *22*(1), 70-79. *https://doi.org/10.7315/CDE.2017.070*

[15] Xie, W. C., Yang, Y. Y., Li, Z. H., Wang, J. W., & Zhang, M. (2018, June). An information hiding algorithm for hevc videos based on pu partitioning modes. In *International Conference on Cloud Computing and Security* (pp. 252-264). Springer, Cham. *https://doi.org/10.1007/978-3-030-00015-8_22*

[16] Miao, Y., Deng, F., Chen, Y., & Guan, H. (2016). Retracted: Detection of volatile organic compounds released by wood furniture based on a cataluminescence test system. *Luminescence*, *31*(2), 407-413. *https://doi.org/10.1002/bio.2974*

[17] Oyabu, T., Kimura, H., & Ishizaka, S. (1995). Indoor air-pollution detector using tin-oxide gas sensor. *Sensors and Materials*, *7*, 431-436. *https://myukk.org/SM2017/article.php?ss=10219*

[18] Oyabu, T., & Kimura, H. (1997). Detection of Relative Gaseous Indoor Air-pollution by Tin Oxide Gas Sensor using Production System. *IEEJ Transactions on Sensors and Micromachines*, *117*(5), 243-249. *10.1541/ieejsmas.117.243*

[19] Wu, F., Jacobs, D., Mitchell, C., Miller, D., & Karol, M. H. (2007). Improving indoor environmental quality for public health: impediments and policy recommendations. *Environmental health perspectives*, *115*(6), 953-957. *https://doi.org/10.1289/ehp.8986*

[20] Bluyssen, P. M., & Cox, C. (2002). Indoor environment quality and upgrading of European office buildings. *Energy and Buildings*, *34*(2), 155-162. *https://doi.org/10.1016/S0378-7788(01)00101-3*

[21] Mirmohammadi, M., Hakimi Ibrahim, M., Ahmad, A., Kadir, M. O. A., Mohammadyan, M., & Mirashrafi, S. B. (2010). Indoor air pollution evaluation with emphasize on HDI and biological assessment of HDA in the polyurethane factories. *Environmental monitoring and assessment*, *165*(1), 341-347. *https://doi.org/10.1007/s10661-009-0950-5*

[22] Dyck, R., Sadiq, R., Rodriguez, M. J., Simard, S., & Tardif, R. (2011). Trihalomethane exposures in indoor swimming pools: a level III fugacity model. *Water research*, *45*(16), 5084-5098. *https://doi.org/10.1016/j.watres.2011.07.005*

[23] Gutiérrez, A. F., Brittle, S., Richardson, T. H., & Dunbar, A. (2014). A proto-type sensor for volatile organic compounds based on magnesium porphyrin molecular films. *Sensors and Actuators B: Chemical*, *202*, 854-860. *https://doi.org/10.1016/j.snb.2014.05.082*

Analysis Platform of Rail Transit Vehicle Signal System Based on Data Mining

Chunying Li*, Zhonghua Mu
[1]Department of Electronic Engineering of Zhengzhou Railway Vocational & Technical College, Zhengzhou, Henan, 451460, China
E-mail: chunyingli7@126.com, zhonghuamu6@163.com
*Corresponding author

Keywords: rail transit, data mining, detection, maintenance, vehicle signal system

Received: January 26, 2022

According to the increasing demand of interactive information of rail transit on-board signal equipment, a rail transit on-board monitoring and maintenance system based on data mining is proposed. The association rules for operation data acquisition are defined and based on these rules a correlation rules algorithm is proposed to obtain more reliable understanding and operation quality evaluation of train operation information. A new approach for the safety analysis is proposed for the analysis of failure mode of rail braking system. The proposed approach uses Bayes method reasoning to introduce reliability and probability analysis of braking system, hence guiding the maintenance strategy of the system. From a lot of logs, quickly find key issues, applied in the train test and repair field. The simulation experiment results show that after analyzing the simulation data and the curve, the system extraction results have certain error in the manual calculation results, and the error value is between 0.5 and 0.6, but the overall meets the actual work needs, and optimize the invalid data to reduce the error. The reliable operation and maintainability of the system are verified.

Povzetek: Razvit je sistem za spremljanje železniškega prometa s podatkovnim rudarjenjem z uporabo Bayesove metode za zanesljivost.

1 Introduction

With the rapid development of rail transit, the degree of automation and the amount of interactive information of on-board signal equipment are increasing, while the maintenance time is continuously shortening, which puts forward higher requirements for the highly reliable operation and safe maintenance of signal equipment. The traditional way of tracking the operation of software and hardware by separately recording error codes in each subunit [1], obviously cannot meet the requirements, and requires an intelligent, concise and user-friendly human-machine interface, which can grasp the overall situation of the vehicle signal equipment in real time. It can provide the expected operation data for testers and maintenance personnel, and realize rapid commissioning and maintenance of trains and lines. At present, in the field of urban rail transit, the application of on-board monitoring and maintenance systems is still relatively rare. In the process of developing the domestic vehicle signal system, the existing achievements of many parties were used for reference, and the multi-source information and data were graded in the data processing. Converted into meaningful data, resulting in abnormal monitoring to assist debugging and maintenance personnel in analyzing equipment failures.

In this paper, based on data mining, the design of rail transit on-board monitoring and maintenance system (TIU, Transit-vehicle Interface Unit) is designed. The main purpose of the design is to comprehensively record the running status of on-board signal equipment), ATO (Automatic Train Operation) and other equipment operation status real-time display and comprehensive analysis, and provide functions such as log download, format conversion and offline analysis, so as to achieve highly reliable operation and safe maintenance of signal equipment. The traffic vehicle monitoring and maintenance system is shown in Figure 1.

In the field of railroad transport, SARM (Security, availability, reliability and maintainability) is generally used to study the functional unwavering quality of gear. In 1986, Sweden gave the primary SARM record necessity in a delicate for the acquisition of fast trains, expecting providers to focus on their unwavering quality, viability and security, and to guarantee that all pointers meet a predefined esteem in the wake of dispatching. During the 1970s, Shinkansen disappointment information was dissected in Japan and enhancements to the plan worked on train unwavering quality. The trains in France have severe SARM prerequisites at the plan stage and their upkeep costs are diminishing step by step. Jing *et al.* played out an assessment of the difficulty free state of SARM on a high velocity rail line [2]. Ma *et al.* brings SARM into metropolitan rail line train checking framework [3]. Liu *et al.* dissected the critical components of SARM control in the rail line industry [4]. Zhong *et al.* present a Bayesian technique for the development of probabilistic organizations and play out an assessment of the framework's credibility [5]. Zhou *et* *al.* subjectively dissect the shortcoming tree of the pressure driven sponsor in the control arrangement of a business vehicle [6]. Wu *et al.* proposed a Bayesian piecewise investigation of a straight model [7]. By changing the information, a Bayesian neighborhood direct model with prescient and nearby straight circulation of boundaries is gotten. Zhong *et al.* dissected the SARM control measures taken at various phases of the flagging framework [8]. Vega *et al.* clear up how for use SARM to control a rail route flagging framework [9]. Horton *et al.* introduced a SARM examination of a rapid rail route power supply framework [10].

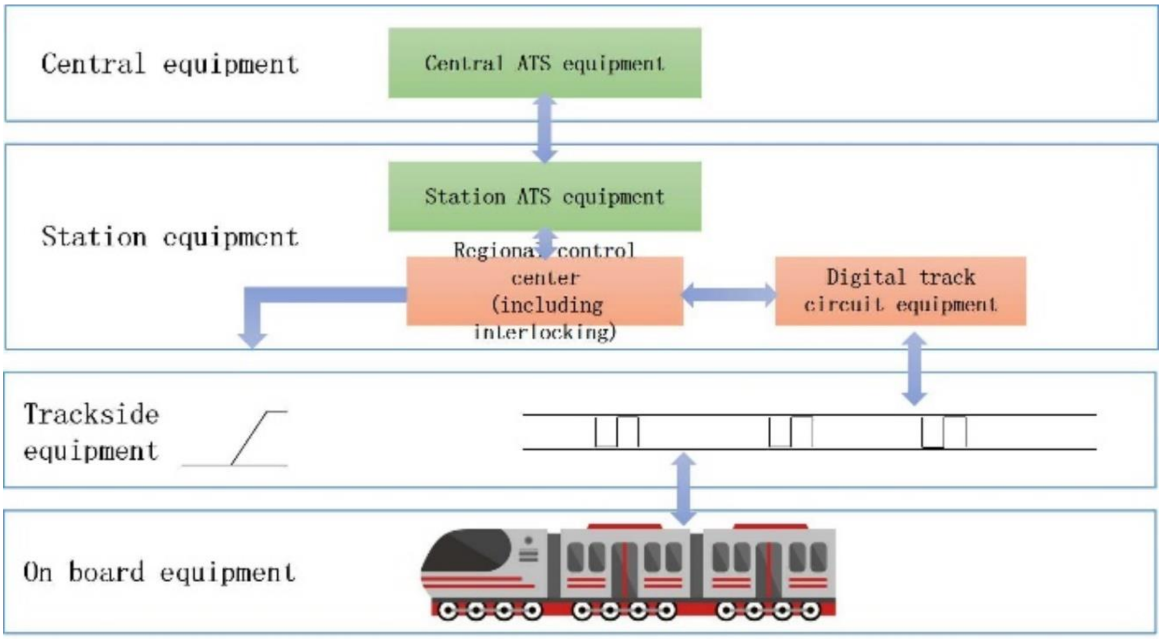


Figure 1: Rail transit vehicle monitoring and maintenance system.

The GO technique is broadly utilized in SARM examination of rail transport frameworks [11]; Zhang *et al.* [12] apply SARM to 9 electromechanical subsystems of travel line 1 of Chengdu Rail, yet there are a few cases effectively utilized in air powered brake frameworks. Numerous specialists utilize the GO-FLOW strategy in rapid rail line security examination [13]. A few specialists apply large information to rail line traffic activity and the board to direct upkeep techniques [14, 15]. The rest of this article is organized as: Section 2 presents the related works in various domains. Section 3 consists of methods comprising the concept. Results and analysis are discussed in Section 4 followed by concluding remarks in section 5.

2 Related work

With the continuous advancement of technology, the technology of video collection data has developed rapidly. Pedestrian movement trajectories can be obtained through video, so as to analyze the characteristics of pedestrian traffic behavior. Zhao *et al.* developed video-based Petrack software for automatic or semi-automatic identification and determination of pedestrian motion positions and trajectories [16]. Wang *et al.* developed a Kinect-based pedestrian trajectory extraction technology for long-term high-precision pedestrian trajectory extraction [17]. Zhigang *et al.* conducted a long-term study on the microscopic traffic behavior of pedestrians. The observed pedestrians need to wear hats of different colors. Using the hats as recognition conditions, the pattern recognition of the captured video images can be used to obtain the motion trajectories of ordinary behaviors and specific behaviors [18]. Chen *et al.* propagandized the pedestrian detection and tracking technology based on Blob analysis, and developed a passenger micro-traffic behavior parameter acquisition system (Ped Trace) to extract traffic behavior characteristic data such as trajectory, passenger speed, pedestrian distance, and acceleration [19]. Ding *et al.* used two-way channel monitoring in urban rail transit stations to manually determine the walking trajectory according to the projected position of the center of gravity of pedestrians' feet, and analyzed the characteristics and laws of pedestrians' overtaking traffic behavior [20]. With the increase of urban population density, the scale of subway construction is also getting larger and larger, and the signal system of rail transit is facing the challenge of more efficient and safer demand. The mobile block train operation control system represented by the Communication Based Train Control System (CBTC) [21] has been popularized to replace the fixed block, and is currently being used for train-to-vehicle communication and fully automatic operation (Fully Automatic Operation, FAO) to further evolve. In order to achieve a shorter running interval of trains in operation, on the one hand, it is necessary to shorten the fault recovery time of the system and make efforts in the direction of immediate maintenance; at the same time, it is also necessary to study the status information of the system and equipment to achieve fault prediction and status repair [22].

Domestic research on PHM (Prognostics Health Management) started later than foreign countries, but relatively speaking, domestic research on PHM in military equipment was earlier [23]. Since the 21st century, the research related to PHM technology has made great progress, mainly reflected in the research of health management and other disciplines. In terms of the top-level design of the PHM system, the aerospace-related research is relatively in-depth [24], and some progress has been made in the PHM research in the fields of machinery, electromechanical, and electronics. The operation safety of the subway has become the research focus of the subway operating companies. In recent years, PHM has gradually been applied to the operation and maintenance management of the subway [25]. However, due to the late start, mature system-level products have not yet appeared. Therefore, in order to improve the operation safety of the subway, improve the maintenance and maintenance efficiency of key equipment, and achieve the purpose of reducing the cost and increasing the efficiency of the subway, the subway system equipment, especially the signal The PHM research of the device is extremely critical [26]. PHM's research in the field of subway is mainly about the health management of high-speed rail equipment, and the application of fault prediction and health management technology to achieve equipment maintenance. For example, algorithms such as VQ and DTW are introduced in the detection of vehicle axle temperature, and the technology originally used in voice signal processing is applied to train health management.

To sum up, based on the current status of maintenance methods of subway equipment at home and abroad and the experience of other industries, combined with the characteristics of domestic rail transit technology, it is proposed to study the maintenance and management technology of rail transit on-board equipment supported by rail transit big data information. development direction [27]. The proposed work can further be extended by using integration approaches of Artificial Intelligence and Machine learning as studied from several studies [28-30]. A technique which considers wavelet frames for micropolar fluid flow is used for high mass transfer [31]. The vibration over the sandwich plates of laminated skew is studied through finite element [32]. The numerical simulation based on space time fractional equation are evaluated [33].

3 System design

In this section, the system structure, functions and association rules based on data mining technique is provided.

3.1 System structure

The system includes a vehicle-mounted unit (lower computer) and a portable maintenance terminal (upper computer), and its hardware and software both adopt a modular design [34]. The on-board unit is placed in the signal cabinet of the train, and has various modes such as network interface, RS-232/422 serial interface, MVB bus interface, etc. It can adapt to the LAN connection method required by ATP, ATO and other equipment, and the required RS-232/422 serial cable connection method, and the MVB bus method used by the vehicle's train management system, etc. The portable maintenance terminal is realized by using a notebook computer. When necessary, it can be connected to the maintenance port of the TIU vehicle-mounted unit by using the RJ-45 network interface. The system structure diagram is shown in Figure 2.

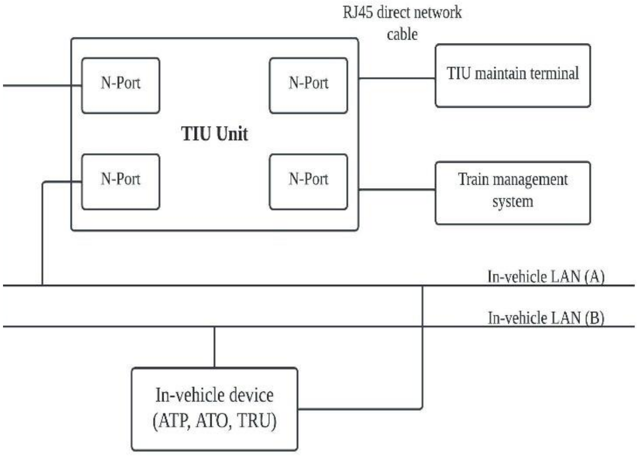


Figure 2: System structure diagram.

A. *Hardware structure*

CPCI bus technology has the technical characteristics of high openness, high reliability, high versatility, hot swap ability, and good anti-vibration and heat dissipation. Therefore, the CPCI bus architecture is adopted in the hardware design of the vehicle-mounted unit. All TIU boards are installed in a 3U cage. It mainly includes power board, CPU motherboard, analog/digital mixed I/O board, Ethernet card, etc. [35].

The power supply board supplies power for the whole system, selects DC110V/±12V/5V/3.3V standard power module, and sets up galvanic isolation device and filter voltage regulation protection circuit to ensure stable and safe voltage to ATO equipment.

i. The power supply board supplies power for the whole system, selects DC110V/±12V/5V/3.3V standard power module, and sets up galvanic isolation device and filter voltage regulation protection circuit to ensure stable and safe voltage to ATO equipment.

ii. The CPU board is the core of computing and control. The TIU needs to exchange data externally, and it needs to output display and input data for the convenience of debugging. Therefore, the CPU board has high reliability with standard serial port, USB interface, VGA interface and external keyboard and mouse.

equipment.

iii. The analog/digital I/O board is the input and output interface of the TIU, so the analog/digital I/O board has enough input and output channels, including digital input and output, analog output and pulse input.

iv. The Ethernet card is the network interface for the communication between the TIU and other on-board components, and a device with a sufficient number of ports and a firm and reliable port connection is selected. All TIU components are designed according to the wide temperature standard of -40℃～+85℃, and the environmental adaptability and electromagnetic compatibility characteristics conform to the relevant technical standards of rail transit on-board equipment.

B. Software structure

Due to the different operating platforms, the software structure includes upper computer and lower computer software. The upper computer software is used for data display and analysis, and is developed using the Windows platform; the lower computer software plays the role of data storage and real-time forwarding, and is developed using a tailored version of Linux. Each software in turn contains multiple modules [36].

The software structure diagram of the lower computer is shown in Figure 3. The software of the lower computer is divided into the bottom general module, the business processing module and the task management module. The low-level general module is closely related to the operating system and hardware drivers. Through the modular programming method, the network port, serial port, file IO, etc. are complicated to set up, and the complex low-level functions are encapsulated into a general module with a simple interface, so as to facilitate the calling of the business processing module. The business processing module separates the processing sub-modules of each data type into a process, and establishes a separate communication channel for it, so as to ensure that the data of different subsystems such as ATP and ATO can be processed in parallel inside the TIU without affecting each other; The task management module is responsible for program startup management, process daemon during program operation, and program exit management to ensure program execution branching and running stability.

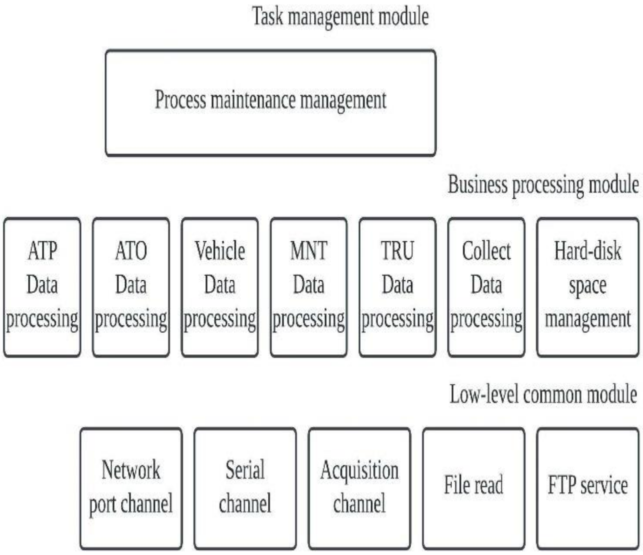


Figure 3: Lower computer software structure diagram.

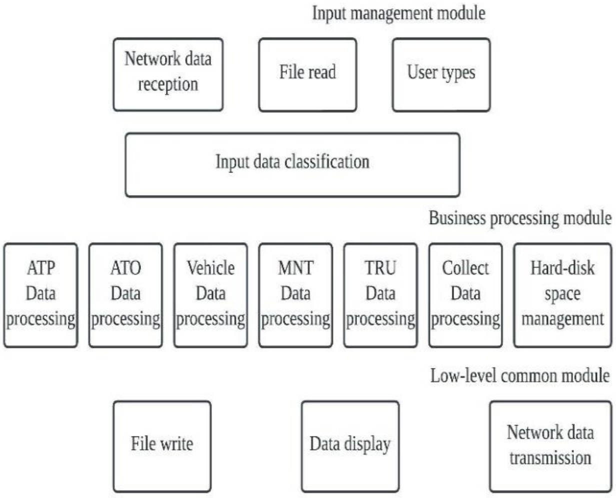


Figure 4: Host computer software structure diagram.

The software structure diagram of the host computer is shown in Figure 4. The upper computer software is divided into input management module, business processing module and output management module. The input management module encapsulates the complex data source processing functions such as network ports, file IO, and user interaction into sub-modules with simple interfaces and convenient calls; the business processing module is responsible for classifying and processing different input data and opening up independent data memory space , establish an independent data processing thread to ensure that the data of different subsystems such as ATP and ATO can be processed synchronously and independently output on the host computer [37]; the output management module is responsible for the module encapsulation of the output function, so as to complete the processing of the business processing module. The data can be classified and displayed, stored and recorded, and forwarded through the network.

3.2 System functions

3.2.1 Lower computer function

The lower computer software is responsible for the acquisition and forwarding of the underlying data, and runs in the on-board TIU host. Real-time forwarding to the maintenance terminal; ① Record the operation data and alarm data of vehicle-mounted signal equipment such as ATP and ATO; ② Relay the operation data of vehicle-mounted signal equipment such as ATP and ATO to the maintenance terminal in real time; ③ transfer the configuration information frame between ATP and MNT; ④ cooperate with MNT to complete TIU's own parameter configuration; ⑤ receive the time synchronization information of ATS and synchronize the local clock; Communication information; ⑦ Collect fan, vehicle power status, and send power-off protection frame to ATO; ⑧ Automatically maintain hard disk space to ensure effective storage of records; ⑨ FTP data service, provide download function of record files.

3.2.2 Host computer configuration

As the window of the system, the host computer software runs in the external maintenance terminal (MNT), the main functions: ① real-time data display, real-time display of the operation and alarm data of ATP, ATO and other equipment; ② offline log display, the display is saved in the machine ③ TIU data download, download the log files stored in the on-board TIU host through FTP; ④ Data format conversion, convert the log file into Excel form; ⑤ System parameter configuration, configuration the system of ATP, ATO, TIU and other systems Parameters; ⑥ Log analysis, assist testers to analyze log data, during the analysis process, the program uses data mining technology to comprehensively analyze logs from different sources and different time periods, so as to select possible useful states or faults [38].

3.3 Application of data mining technology in log analysis

3.3.1 Mining algorithm selection and introduction

The data recorded by TIU has the characteristics of being complex, multi-level and uncertain: ① There are various sources, including the status of on-board equipment such as ATP and ATO, as well as the information of vehicles and trackside equipment; ② There are many states, including door status and speed information , location information, alarm information and other thousands of information states; ③ There are various types of states, and the data types of different information states vary widely, such as Boolean, integer, floating point, string, etc.; ④ The dispersion of log records, The randomness of the time span, etc. According to these characteristics, a one-dimensional simple association rule model that is simple in form, easy to understand, and can effectively capture the relationship between data is adopted in the log analysis algorithm [39].

3.3.2 Principles of association rules

Association rule mining can be formulated as follows: Let I = (i1, i2, ...in) be a set of items, and T = (t1, t2, ...tn) a set of transactions, where each transaction ti is a set of items and satisfies ti. An association rule for ∈I: X→Y, where X∈I, Y∈I, and X∩Y=. Support and confidence are two commonly used indicators to measure the strength of association rules. The support of rule X→Y refers to the percentage of transactions that contain itemset X∪Y in transaction set T, so the support of rule Ts represents the frequency of rule use in transaction set T [40].

$$T_s = \frac{(X \bigcup Y)\cdot count}{n} \tag{1}$$

where n is the total number of transactions in T. Confidence rule, the confidence of X→Y refers to the percentage of transactions that contain both X and Y to all transactions that contain X. It can be regarded as an estimate of the conditional probability P(X|Y). The confidence Hs determines the predictability of the rule [41].

$$H_s = \frac{(X \bigcup Y)\cdot count}{X \cdot count} \tag{2}$$

Association rule mining refers to finding out the association rules in T whose support and confidence are higher than a user-specified minimum support (min-supp) and minimum confidence (minconf) respectively [42].

3.3.3 Algorithms of association rules

The algorithm of association rules, the Apriori algorithm that uses candidate item sets to find frequent item sets, is mainly divided into two steps: 1. Generate all frequent item sets, one frequent itemset is an itemset whose support is higher than minsup; 2. From the frequent item sets Generate all trusted association rules, a trusted association rule is a rule with a confidence greater than minconf. Association rules only need to be generated based on frequent item sets [43]: extracting all association rules from frequent item sets f needs to use all non-empty subsets of f, let a be any non-empty subset of f, then: (f-a) →a, if the confidence Ts satisfies:

$$T_s = \frac{f \cdot count}{(f-a)\cdot count} \ge \min conf \qquad (3)$$

An association rule generation algorithm similar to frequent itemset generation can be used: first generate 1-association rules with only one item of all consequences from the frequent itemset f (k-itemset is a set containing k items), and then use the association rule. The rules generate consequent 2-association rules, which are recursive in turn to generate all frequency sets.

4 Results and analysis

Through the design method described in the previous chapter, in the log-assisted analysis, the association rules in data mining can be applied as follows.

The first step is to establish a model, by setting the granularity of conditions and conclusions, artificially setting intervals, selecting single values, setting fuzzy values, and setting the certainty of the rules, mainly in the setting of precise rules and conceptual rules, The train operation data is used as the training data set, and the classification data is tested with the above settings. This process can remove redundant data and irrelevant attributes, and find all high-frequency item groups. For example, if you need to count the data of train stops, you need to set the query time period, a certain platform number, the stop sign, and whether the distance from the stop sign is less than the set value. The above settings are a group of high-frequency items, while in another group in the project group, such as the door opening process, there are also check stop signs and platform numbers.

The second step is to find the support and confidence of the high-frequency item group, set the thresholds for both, and discover the association rules. As shown in the above example, if the stop sign appears in both project groups, the support degree is 0.5, and the platform number also appears in the two projects at the same time. When the stop sign appears, the confidence level is 1. Confidence is always greater than support. If the support degree is greater than the user-set value (temporarily set to 0.5), then these two items are frequent item sets.

The third step is to display and evaluate association rules. Simple one-dimensional association rules are used to calculate frequent item sets, and strong association rules are generated from frequency sets, which must satisfy both minimum support and minimum confidence (temporarily set to 0.5), thus generating two valid values, and so on. The item set is calculated, all frequency sets are generated, and the calculation results are displayed, and finally a subjective screening is carried out. Obvious irrelevant information is excluded. In this way, the association rule mining of a set of information is completed. By mining different information and applying association rules for different purposes, possible errors of programs and possible failures of equipment can be screened out [44]. The abnormal information screened out when the transponder fails is shown in Table 1, which means that the transponder receives the default message at a certain speed and at a certain time. The abnormality will not affect the train running in continuous mode. Under the formula, the abnormality becomes a dominant fault [45].

Table 1: Abnormal information filtered out when the transponder fails

Line number	Time	Speed	Transponder message	Transponder	Screening amount
2711	6:22:54	70.99	default	0x012E004A	1
2712	6:22:55	70.93	default	0x012E004A	1
2713	6:22:55	70.96	default	0x012E004A	1
2714	6:22:56	71.04	default	0x012E004A	1
2715	6:22:57	70.96	default	0x012E004A	1
2716	6:22:57	70.74	default	0x012E004A	1
2717	6:22:58	70.53	default	0x012E004A	1
2718	6:22:58	70.74	default	0x012E004A	1
2719	6:22:59	70.58	default	0x012E004A	1
2720	6:22:59	70.68	default	0x012E004A	1

By monitoring such abnormal signs, it is possible to regularly check the running status of the backup system equipment, and find the fault of the backup system equipment in time. To avoid the situation that when the main system is working, the backup system has failed but has not been repaired in time, and the backup system cannot take over in time after the failure of the main system [46].

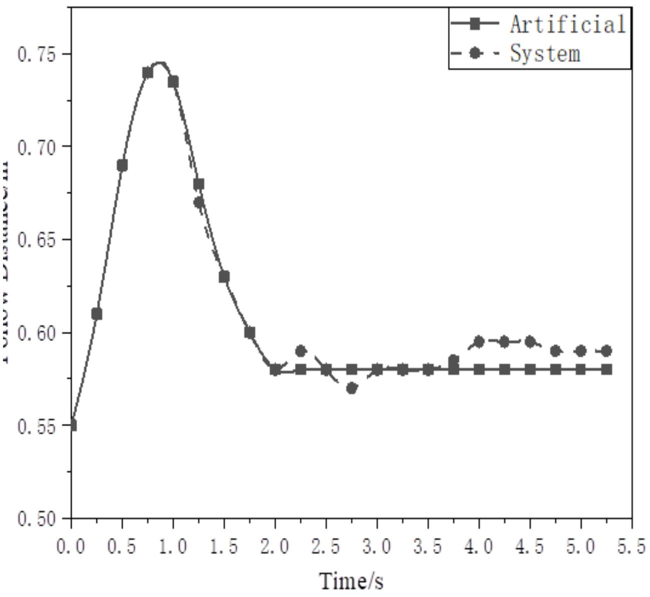


Figure 5: Following distance feature extraction results.

Use the designed rail transit on-board signal system to conduct simulation experiments to extract various behavioral characteristic parameters, and compare the results extracted by the system with the manual calculation results, as shown in Figures 5, 6, and 7 below. After analyzing the simulation data and curves, it can be seen that there is a certain error in the extraction results of the system compared with the manual calculation results, error values are between 0.5 and 0.6, but it generally meets the actual work requirements. At the same time, the system data mining method can be further optimized. Invalid data such as inflection point data are cleaned to reduce errors.

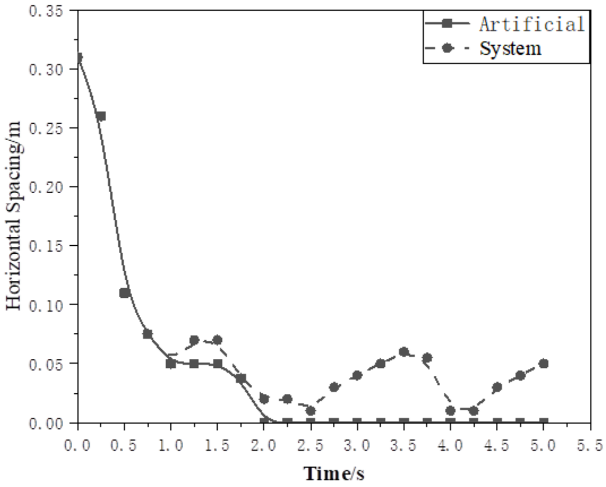


Figure 6: Horizontal spacing feature extraction results.

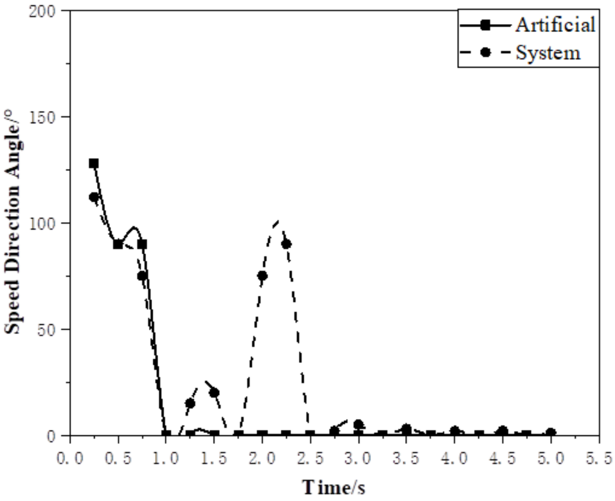


Figure 7: Feature extraction results of included angle in velocity direction.

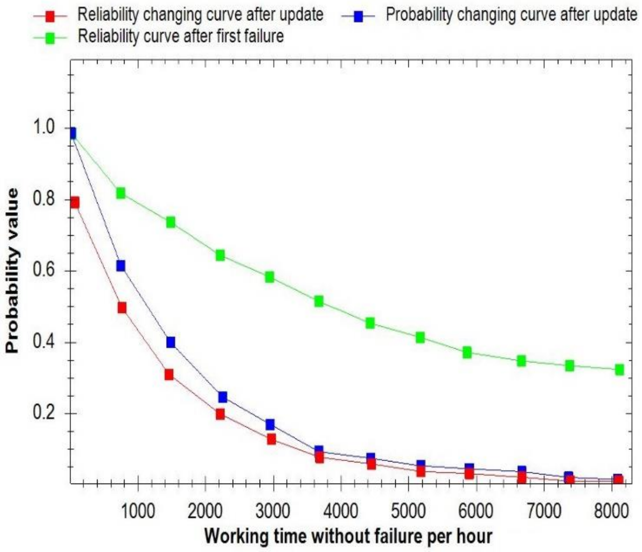


Figure 8: Results through simulation.

Figure 8 depicts the simulation outcomes of the proposed system. The reliability graph before the primary failure is equivalent to the typical probability graph, however it varies from one another after the updating of failure and maintenance. The pattern of the typical probability graph after the framework update is more awful than before the primary failure. This is on the grounds that a few old units stay in their ordinary states in the refreshed framework, making it simpler for the framework to go into a failure state. The measured reliability is additionally lower. After the framework is refreshed, the dependability is more awful than the first framework, albeit the framework can work typically. Over the long haul, the probability of typical activity moves toward the dependability of the framework. It ought to be noticed that unwavering quality is the premise of security, while safety mirrors the continuous condition of dependability.

5 Conclusions

By designing a rail transit on-board system platform based on data mining, this paper analyzes the system structure and functions required in the process of rail transit on-board monitoring and maintenance, and briefly introduces the data mining algorithm used in the log-assisted analysis process. Finally, the reliability and maintainability of the system are verified by simulation experiments. The application of this system not only realizes the centralized recording of the operation status and alarm of the vehicle signal system, but also provides a powerful tool for the debugging personnel to monitor the operation of the equipment and analyze the program loopholes, and greatly improve the efficiency of the maintenance personnel to analyze the logs. And make fault early warning a possibility to prevent problems before they happen.

References

[1] Huang, C., & Huang, Y. (2021). Urban rail transit signal and control based on Internet of Things. *Journal of High Speed Networks*, (Preprint), 1-14. *10.3233/JHS-210664*

[2] Jing, G., Ding, D., & Liu, X. (2019). High-speed railway ballast flight mechanism analysis and risk management–A literature review. *Construction and Building Materials*, *223*, 629-642. https://doi.org/10.1016/j.conbuildmat.2019.06.194

[3] Ma, J., Su, C., Yang, Y., Wu, M., & Jiang, B. (2016). The field test for influence of ram-compacted piles with bearing base on settlement of embankments in China Beijing-Shanghai high-speed railway on deep soft soil. *Japanese Geotechnical Society Special Publication*, *2*(3), 217-220. https://doi.org/10.3208/jgssp.CHN-49

[4] Liu, Z., Wang, L., Li, C., & Han, Z. (2017). A high-precision loose strands diagnosis approach for isoelectric line in high-speed railway. *IEEE Transactions on Industrial Informatics*, *14*(3), 1067-1077. 10.1109/TII.2017.2774242

[5] Zhong, J., Liu, Z., Han, Z., Han, Y., & Zhang, W. (2018). A CNN-based defect inspection method for catenary split pins in high-speed railway. *IEEE Transactions on Instrumentation and Measurement*, *68*(8), 2849-2860. 10.1109/TIM.2018.2871353

[6] Zhou, L., Tong, L. C., Chen, J., Tang, J., & Zhou, X. (2017). Joint optimization of high-speed train timetables and speed profiles: A unified modeling approach using space-time-speed grid networks. *Transportation Research Part B: Methodological*, *97*, 157-181. https://doi.org/10.1016/j.trb.2017.01.002

[7] Wu, Y., Qin, Y., Qian, Y., & Guo, F. (2021). Automatic detection of arbitrarily oriented fastener defect in high-speed railway. *Automation in Construction*, *131*, 103913. https://doi.org/10.1016/j.autcon.2021.103913

[8] Zhong, Z. D., Ai, B., Zhu, G., Wu, H., Xiong, L., Wang, F. G., & He, R. S. (2018). *Dedicated mobile communications for high-speed railway* (Vol. 22). Heidelberg: Springer. https://doi.org/10.1007/978-3-662-54860-8

[9] Vega, J., Fraile, A., Alarcon, E., & Hermanns, L. (2012). Dynamic response of underpasses for high-speed train lines. *Journal of Sound and Vibration*, *331*(23), 5125-5140. https://doi.org/10.1016/j.jsv.2012.07.005

[10] Horton, M., Connolly, D. P., & Yu, Z. (2017). Rail trackbed and performance testing of stabilised sub-ballast in normal and high-speed environments. *Procedia engineering*, *189*, 924-931. https://doi.org/10.1016/j.proeng.2017.05.143

[11] Zhan, S., Wang, P., Wong, S. C., & Lo, S. M. (2022). Energy-efficient high-speed train rescheduling during a major disruption. *Transportation Research Part E: Logistics and Transportation Review*, *157*, 102492. https://doi.org/10.1016/j.tre.2021.102492

[12] Zhang, H., Peng, Y., Hou, L., Wang, D., Tian, G., & Li, Z. (2019). Multistage impact energy distribution for whole vehicles in high-speed train collisions: modeling and solution methodology. *IEEE Transactions on Industrial Informatics*, *16*(4), 2486-2499. 10.1109/TII.2019.2936048

[13] Xu, P., Corman, F., Peng, Q., & Luan, X. (2017). A timetable rescheduling approach and transition phases for high-speed railway traffic during disruptions. *Transportation Research Record*, *2607*(1), 82-92. https://doi.org/10.3141/2607-11

[14] Yang, S., Song, K., & Zhu, G. (2019). Stochastic process and simulation of traction load for high speed railways. *IEEE Access*, *7*, 76049-76060. 10.1109/ACCESS.2019.2921093

[15] Mao, Q., Cui, H., Hu, Q., & Ren, X. (2018). A rigorous fastener inspection approach for high-speed railway from structured light sensors. *ISPRS Journal of Photogrammetry and Remote Sensing*, *143*, 249-267. https://doi.org/10.1016/j.isprsjprs.2017.11.007

[16] Zhao, J., Ma, C., Xiao, X., & Jiang, Y. (2022). Research on deformation law of guide rails caused by mine vertical shafts under non-mining action. *Engineering Failure Analysis*, *134*, 106089. *https://doi.org/10.1016/j.engfailanal.2022.106089*

[17] Wang, K., Yan, X., Yuan, Y., Jiang, X., Lodewijks, G., & Negenborn, R. R. (2017, August). Study on route division for ship energy efficiency optimization based on big environment data. In *2017 4th International Conference on Transportation Information and Safety (ICTIS)* (pp. 111-116). IEEE. *10.1109/ICTIS.2017.8047752*

[18] Zhigang, M. (2014, October). Grey Prediction of Urban Rail Transit Machine-Electric Equipment Fault Based on Data Mining. In *2014 7th International Conference on Intelligent Computation Technology and Automation* (pp. 284-287). IEEE. *10.1109/ICICTA.2014.76*

[19] Chen, X., Guo, Y., Li, B., Ge, M., & Xu, C. (2015). Analysis of Dynamic Passenger Flow in Urban Rail Transit Based on Data Mining. In *ICTE 2015* (pp. 2035-2042). *https://doi.org/10.1061/9780784479384.259*

[20] Ding, X., Yang, X., Hu, H., & Liu, Z. (2017). The safety management of urban rail transit based on operation fault log. *Safety science*, *94*, 10-16. *https://doi.org/10.1016/j.ssci.2016.12.015*

[21] Ming, Z. (2015, September). Decision approach of maintenance for urban rail transit based on equipment supervision data mining. In *2015 IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)* (Vol. 1, pp. 376-380). IEEE. *10.1109/IDAACS.2015.7340761*

[22] Chen, G. (2018, January). Optimization Design of Passenger Flow in Rail Transit Station in Shanghai Based on Data Mining. In *2017 4th International Conference on Machinery, Materials and Computer (MACMC 2017)* (pp. 583-588). Atlantis Press. *https://doi.org/10.2991/macmc-17.2018.108*

[23] Bai, L., Wang, F. Z., & Zhang, M. (2015). Application of geographic information system (gis) in urban rail transit construction safety and operation monitoring. In *Applied Mechanics and Materials* (Vol. 743, pp. 692-697). Trans Tech Publications Ltd. *https://doi.org/10.4028/www.scientific.net/AMM.743.692*

[24] Liu, X., & Yang, X. (2019, March). Identifying technological innovation capability of high-speed rail industry based on patent analysis. In *2019 8th International Conference on Industrial Technology and Management (ICITM)* (pp. 127-131). IEEE. *10.1109/ICITM.2019.8710710*

[25] Zhang, S. (2012). Data Mining and Analysis of Integrated Circuit Card Data in Subway. In *Future Wireless Networks and Information Systems* (pp. 739-744). Springer, Berlin, Heidelberg. *https://doi.org/10.1007/978-3-642-27326-1_95*

[26] Fu, L., Wang, X., Zhao, H., & Li, M. (2022). Interactions among safety risks in metro deep foundation pit projects: An association rule mining-based modeling framework. *Reliability Engineering & System Safety*, *221*, 108381. *https://doi.org/10.1016/j.ress.2022.108381*

[27] Zimbalist, A. (2013). Inequality in intercollegiate athletics: Origins, trends and policies. *Journal of Intercollegiate Sport*, *6*(1), 5-24. *https://doi.org/10.1123/jis.6.1.5*

[28] Wang, H., Hao, L., Sharma, A., & Kukkar, A. (2022). Automatic control of computer application data processing system based on artificial intelligence. *Journal of Intelligent Systems*, *31*(1), 177-192. *https://doi.org/10.1515/jisys-2022-0007*

[29] Sun, L., Gupta, R. K., & Sharma, A. (2022). Review and potential for artificial intelligence in healthcare. *International Journal of System Assurance Engineering and Management*, *13*(1), 54-62. *https://doi.org/10.1007/s13198-021-01221-9*

[30] Cai, Y., & Sharma, A. (2021). Swarm intelligence optimization: an exploration and application of machine learning technology. *Journal of Intelligent Systems*, *30*(1), 460-469. *https://doi.org/10.1515/jisys-2020-0084*

[31] Kumbinarasaiah, S., & Raghunatha, K. R. (2021). A novel approach on micropolar fluid flow in a porous channel with high mass transfer via wavelet frames. *Nonlinear Engineering*, *10*(1), 39-45. *https://doi.org/10.1515/nleng-2021-0004*

[32] Dhotre, P. K., & Srinivasa, C. V. (2021). On free vibration of laminated skew sandwich plates: A finite element analysis. *Nonlinear Engineering*, *10*(1), 66-76. *https://doi.org/10.1515/nleng-2021-0006*

[33] Pedram, L., & Rostamy, D. (2021). Numerical simulations of stochastic conformable space–time fractional Kortewegde Vries and Benjamin–Bona–Mahony equations. *Nonlinear Engineering*, *10*(1), 77-90. *https://doi.org/10.1515/nleng-2021-0007*

[34] Duan, Y. Q., Fan, X. Y., Liu, J. C., & Hou, Q. H. (2020). Operating efficiency-based data mining on intensive land use in smart city. *IEEE Access*, *8*, 17253-17262. *10.1109/ACCESS.2020.2967437*

[35] Wang, D. L., Yao, E. J., Yang, Y., & Zhang, Y. S. (2014). Modeling passenger flow distribution based on disaggregate model for urban rail transit. In *Foundations and Practical Applications of Cognitive Systems and Information Processing* (pp. 715-723). Springer, Berlin, Heidelberg. *https://doi.org/10.1007/978-3-642-37835-5_62*

[36] Wang, F., Xu, T., Tang, T., Zhou, M., & Wang, H. (2016). Bilevel feature extraction-based text mining for fault diagnosis of railway systems. *IEEE transactions on intelligent transportation systems*, *18*(1), 49-58. *10.1109/TITS.2016.2521866*

[37] Massa, S., & Puliafito, P. P. (1999, September). An application of data mining to the problem of the university students' dropout using markov chains. In *European conference on principles of data mining and knowledge discovery* (pp. 51-60). Springer, Berlin, Heidelberg. *https://doi.org/10.1007/978-3-540-48247-5_6*

[38] Dai, X., Qiu, H., & Sun, L. (2021). A Data-Efficient Approach for Evacuation Demand Generation and Dissipation Prediction in Urban Rail Transit System. *Sustainability*, *13*(17), 9692. *https://doi.org/10.3390/su13179692*

[39] Tang, Y. (2021). Risk Chain and Key Hazard Management for Urban Rail Transit System

Operation Based on Big Data Mining. *Discrete Dynamics in Nature and Society*, *2021*. *https://doi.org/10.1155/2021/3692151*

[40] Hu, P., Duan, K., Huo, W., Zhang, Q., Zhou, M., & Ba, Y. (2020, November). Design of state Grid shopping mall heating technology application assistant decision system based on big data analysis. In *2020 International Conference on Computer Engineering and Intelligent Control (ICCEIC)* (pp. 97-100). IEEE. *10.1109/ICCEIC51584.2020.00027*

[41] Ming, Z., Xiaofei, W., & Li, B. (2014, June). The Fault Data Mining of Supervision Equipment of Urban Rail Transit Based on Clustering. In *2014 Fifth International Conference on Intelligent Systems Design and Engineering Applications* (pp. 1045-1048). IEEE. *10.1109/ISDEA.2014.230*

[42] Gusikhin, O., Rychtyckyj, N., & Filev, D. (2007). Intelligent systems in the automotive industry: applications and trends. *Knowledge and Information Systems*, *12*(2), 147-168. *https://doi.org/10.1007/s10115-006-0063-1*

[43] Cheng, X., Huang, K., Qu, L., & Li, L. (2020). A cooperative data mining approach for potential urban rail transit demand using probe vehicle trajectories. *IEEE Access*, *8*, 24847-24861. *10.1109/ACCESS.2020.2970863*

[44] Ma, X. L., Wang, Y. H., Chen, F., & Liu, J. F. (2012). Transit smart card data mining for passenger origin information extraction. *Journal of Zhejiang University Science C*, *13*(10), 750-760. *https://doi.org/10.1631/jzus.C12a0049*

[45] Shuai, L., Limin, J., Yong, Q., Bo, Y., & Yanhui, W. (2014). Research on urban rail train passenger door system fault diagnosis using PCA and rough set. *The Open Mechanical Engineering Journal*, *8*(1). *10.2174/1874155X01408010340*

[46] Du, G., Zhang, X., & Ni, S. (2018, October). Discussion on the application of big data in rail transit organization. In *International Conference on Smart Vehicular Technology, Transportation, Communication and Applications* (pp. 312-318). Springer, Cham. *https://doi.org/10.1007/978-3-030-04582-1_36*

Personalized Recommendation System of E-learning Resources Based on Bayesian Classification Algorithm

Xiuhui Wang
Shanxi Datong University, Institute of education science and technology, Datong, Shanxi,037009, China
E-mail: Xiuhuiwang6@163.com

Keywords: learning resources, personalized recommendation, bayesian classification algorithm, apriori algorithm, historical data pruning

Received: February 2, 2022

This article addresses the problem of learners' information trek as well as overload and meet the learners' personalized learning needs by realizing learners' personalized development. In this article, the development scheme of e-learning resources personalized recommendation system based on Bayesian algorithm is proposed. This paper studies the personalized Association recommendation model integrating association rule mining and Bayesian network, thereby improving the association rule mining algorithm by combining historical record pruning and Bayesian network verification. In the process of association rule mining, the proposed methodology is combined with user history and the frequent item sets in association rules are filtered. The item sets below the given threshold are pruned. The pruned item set is input into the Bayesian verification network for personalized verification, and the verification results are sorted and recommended according to the ranking. This is done to give priority to the readers who really like the books. The recommendation model solves the problem of weak personalization in the existing recommendation system to a certain extent. The experiments show that Bayesian network can improve the personalization of association recommendation.

Povzetek: Članek obravnava personalizacijo e-učenja z uporabo Bayesovega algoritma in učenjem asociativnih pravil, ki izboljšuje personalizacijo priporočil in povečuje učinkovitost.

1 Introduction

Recommendation system is a new information discovery mode. It models users' interests and preferences by analyzing users' historical behavior information, and recommends items conforming to their preferences to users according to users' preferences. In this process, users do not need to provide any demand information, and the recommendation system actively pushes information to users [1]. The emergence of the recommendation system solves the problem of how to select items from a large amount of item information when the user's needs are not clear or cannot accurately describe the needs. The emergence of the recommendation system also transforms the way of information acquisition from data search to higher-level information discovery. After the emergence of recommendation system, it has gradually proved to be an effective tool to solve the problem of information overload. In essence, the recommendation system solves the problem of information overload by recommending new items unfamiliar to the user, and the recommended new items are likely to be related to the needs of the user. For each request of the user, the recommendation system uses different recommendation methods and the environment and needs of the user at that time to generate recommendation results according to the user, available items, user historical transaction data and various types of additional information stored in the system. Users may or may not accept these recommended items, and may give explicit or implicit feedback immediately, or give feedback over a period of time [2].

All these user behaviors and feedback information are stored in the recommendation system and become the data source of recommendation when users interact with the recommendation system in the future. With the advent of the education informatization 2.0 era and the rapid development of education big data, learning analysis, artificial intelligence and other technologies, the educational form will change profoundly, and promoting students' personalized development has become the core demand. The realization of students' personalized development is inseparable from the support of personalized learning. However, education is still dominated by the traditional class teaching system. Due to the large number of students, it is difficult for teachers to "teach students according to their aptitude" with their personal ability. As a result, it is difficult to meet learners' personalized learning needs and hinder students' personalized development. At the same time, the explosive growth of e-learning resources also brings the problem of information Trek and overload to learners, which hinders learners from accurately positioning their own learning resources. Relevant studies believe that the intelligent service of personalized recommendation of e-learning resources can not only solve the problem of students' personalized needs, but also effectively avoid information loss and overload [3]. Therefore, the related research has been widely concerned by scholars in the field of education. Therefore, from the perspective of academic research or practical application, the research on personalized recommendation of e-learning resources is of great significance to educational development. It is an effective method and key way to support learners' personalized development and solve the problem of information Trek and overload. Figure 1 is the conceptual diagram of personalized recommendation system.

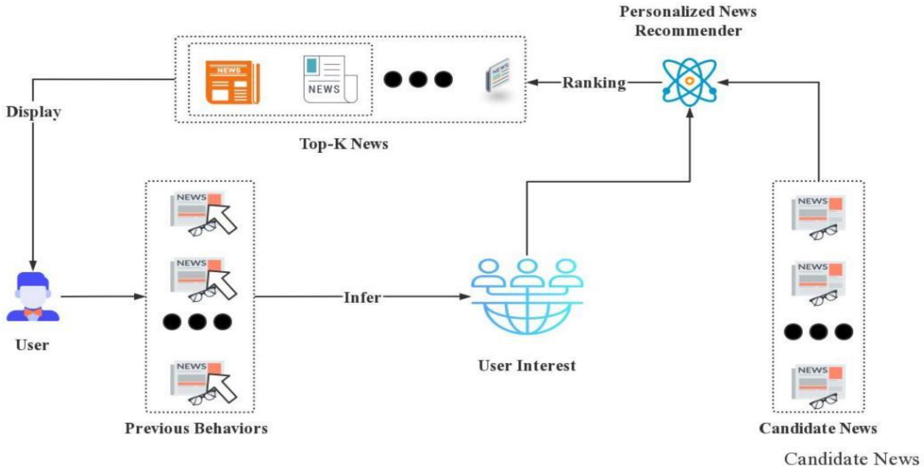


Figure 1: Personalized recommendation system.

This article contributes in the development scheme of e-learning resources personalized recommendation system based on Bayesian algorithm. The personalized Association recommendation model is studies which integrates association rule mining and Bayesian network, thereby improving the association rule mining algorithm by combining historical record pruning and Bayesian network verification. In the process of association rule mining, the proposed methodology is combined with user history and the frequent item sets in association rules are filtered. The item sets below the given threshold are pruned. The pruned item set is input into the Bayesian verification network for personalized verification, and the verification results are sorted and recommended according to the ranking. The recommendation model solves the problem of weak personalization in the existing recommendation system.

The rest of this article is organized as: Literature is presented in section 2 followed by the methodology section discussing the personalized association recommendation model based on Bayesian network in section 3. Results are analyzed in section 4 followed by conclusion in section 5.

2 Related work

In this section various recent work done in the field of recommendation system of e-learning related studies are explored and discussed.

At present, there are many theories and research methods on recommendation system. Rawat and Dwivedi studied the recommendation system based on association rules, compared a series of products purchased by current customers with a series of products purchased by other customers, selected the intersection of products purchased by customers and the product set purchased by current customers, and presented them to customers as recommended products [4]. Zhong proposed a personalized information filtering recommendation method based on simple Bayesian classifier to smooth user rating, which can alleviate sparsity and improve the accuracy of searching nearest neighbors [5]. Benhamdi, *et al.* others proposed a Bayesian network classification algorithm bncar based on association rules. The algorithm uses the association rule mining algorithm to extract the initial candidate edges, and obtains a better Bayesian network structure through the greedy algorithm. Bncar can obtain higher classification performance. Wang and others proposed an eye movement trajectory semantic extraction algorithm by collecting the eye movement parameters of customers observing an object and referring to the idea of find-s algorithm. The algorithm maximizes the distance between positive and negative examples by learning a priori knowledge, determines the weight of eye movement parameters, including gaze time, pupil size, blink times and look back times, uses sebet algorithm to judge whether customers like a commodity according to the distance, and realizes semantic extraction from eye movement trajectory. The study did not consider personalized relevance recommendation [6].

Chao *et al.* proposed an association rule extraction method using improved genetic algorithm, gave a specific algorithm, applied this knowledge to students' teaching management, and made corresponding adjustments to the original teaching system and plan. The algorithm has certain application value in promoting student training and education. This research fails to deeply study the user personalized semantics [7]. Kumar *et al.* proposed ar-sem algorithm based on association rules. Firstly, the algorithm uses association rules to analyze the causal relationship between variables, and combines it with the initial prior knowledge and the opinions of domain experts to further remove meaningless rules and form a knowledge base. Finally, the knowledge base is combined with SEM algorithm to construct Bayesian network. Although this method can improve the accuracy of Bayesian network structure learning to a certain extent, there is no personalized test on the mined association rules [8]. In terms of application, Sharma and Suryavanshi established a Bayesian network model of Anabaena bloom in dams. In the model, the monitoring data is stored in a unified database. By "learning" the relationship probability between factors, such as nutrient load, nutrient concentration of lake water body and fishy concentration, it can be convenient for nonprofessional modelers to use, so as to significantly reduce water treatment costs and operating expenses [9]. Zhang *et al.* put forward the concept of interest degree of attribute set through massive data and data flow analysis, studied the characteristic difference of attribute set as data, and deduced an accurate algorithm to find that the interest degree of all attribute sets exceeds the given threshold. The algorithm finds the most interesting attribute set by specifying similarity and confidence probability [10].

Chinna Gounder Dhanajayan, summarized the research of Bayesian network reasoning algorithms and their development and function expansion in recent 30 years, and compared them from the aspects of complexity, applicability and accuracy. The key links of each algorithm are pointed out, the application of Bayesian network in the field of engineering technology is analyzed and reviewed, and the shortcomings and future research trend of BN are summarized and prospected [11]. Aiming at the problem that traditional association rule mining cannot reflect the semantic measurement between items, Zhu *et al.* studied how to use the utility confidence framework to find association rules, studied a dense representation of mining all minimum antecedent and maximum antecedent association rules, and realized it by using closed itemset (HUCI) and its generator efficiently. In terms of personalized Preference Research [12], Zhang *et al.* use ontology to organize user and service information, and find the content they are interested in according to the user's preference [13]. One similar study that uses wavelet frames for measuring micropolar fluid flow is used for high mass transfer and some other relevant studies also studied that uses various approaches for measuring vibrations, space time fraction [14-16]. This work can be considered for future development by using integration approaches of Artificial Intelligence and Machine learning as studied from several studies [17-19].

3 Research methodology

This section includes the discussion of personalized association recommendation model on the basis of Bayesian network.

3.1 Personalized association recommendation model based on bayesian network

The recommendation model established by association rule mining, historical data pruning and Bayesian network verification is shown in Figure 2. The model includes four functional modules:

1. Module A is the association rule mining module, and the algorithm used in this module is Apriori algorithm.
2. Module B is a history pruning module, which prunes the of association rules with history data, and the item set lower than the given threshold is pruned.
3. Module C is a Bayesian network verification module, which uses Bayesian network to verify the semantics of association rules and sort the associated item sets according to probability priority.
4. Module D is the recommendation strategy formulation module, which formulates the recommendation strategy according to the output results of Bayesian verification network [20]. Apriori algorithm is a frequent itemset algorithm for mining association rules.

The algorithm is divided into two steps: the first step is to retrieve all frequent itemset in the transaction database through iteration, that is, itemset with support not lower than the threshold set by the user; The second step uses frequent itemset to construct rules that meet the minimum trust of users [21]. The specific method is: first, find out the frequent 1-itemset and record it as L_1; Then, L_1 is used to generate candidate itemset C_2, the items in C_2 are determined, and L_2, that is, frequent 2-itemset, is mined; This cycle continues until no more frequent k-itemset can be found. Bayesian network is a probabilistic network. It is a graphical network based on probabilistic reasoning. It obtains other probabilistic information through the information of some variables to solve the uncertainty and incompleteness of some facts in application. It has been widely used in many fields, and Bayesian theorem is the basis of Bayesian network [22]. Bayesian theorem is used to describe the relationship between two conditional probabilities, such as $P(A|B)$ and $P(B|A)$. According to the multiplication rule:

$$P(A \cap B) = P(A) * P(B \mid A) = P(B) * P(A \mid B) \tag{1}$$

Bayesian formula can be derived:

$$P(B \mid A) = P(A \mid B) * P(B) / P(A) \tag{2}$$

This formula is generally generalized to obtain the general Bayesian formula [23]:

$$P(A_i \mid B) = \frac{P(B \mid A)P(A_i)}{\sum_{i=1}^{n} P(B \mid A_i)P(A_i)} \qquad (3)$$

Where, $A_1, \ldots, A_n$ is the complete event group [24], that is: $U_{i=1}^{n} A_i = \Omega, A_j = \varphi, P(A_i) > 0$.

Taking e-book borrowing as an example, D_1 is the borrowing record of readers. The algorithm mines association rules from D_1, and then prunes the mined frequent large itemsets by using the user's historical information [25]. The pruning results are verified by Bayesian network to obtain the verified frequent large itemset, so as to formulate the recommendation strategy. In the following algorithm, algorithm 1 calls algorithm 2 [26].

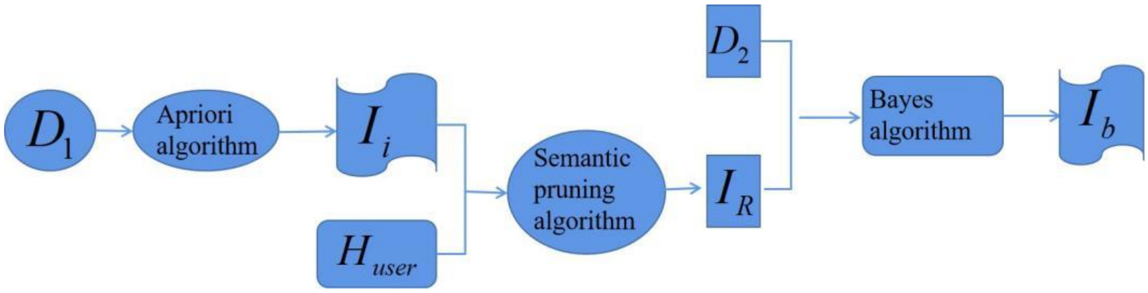


Figure 2: Model diagram

Algorithm 1 Bayesian personalized Association recommendation algorithm

```
Input: $D_1$;
Output: $I_b$;
Begin
1. Sort out readers' borrowing records and get $D_1$.
2. Set the support according to the Apriori algorithm_Degree, generate k itemsets, and calculate the large itemset $I_i, I_x \in I_i, i, k \in [1, n]$to recommend content to user $R_i$, $n$ is the number of items [27].
3. The historical data of the user is recorded as $H_{user}$. combined with $H_{user,}prue(I_i, H_{user})$ is called to perform personalized semantic pruning on $I_i$ and output $I_R = \{I_r | I_r \in I_i\}$.
4. If $I_R$ meets the set $k$ itemset requirements, execute (5), otherwise execute (2) until the requirements are met.
5. Build $D_2$, which is a borrowing record database that can reflect users' preferences.
6. Take $I_R$ as the input of $N$, run Bayes algorithm, and output $I_b$ and $I_b = \{I_y \middle| I_y \in I_i\}$ sorted by readers' preference.
7. Return$(I_b)$; $I_b$ is the content to be recommended to the user [28].
End
```

In algorithm 1, when building D_2, it is necessary to ensure the scientificity of D_2. each record contains the basic information of the user as set u. The basic information of the user includes gender, age and other contents. The basic information of the user needs to be determined according to the actual situation. This algorithm prunes the large itemset mined by using the historical records to delete the records that are not in good agreement with the historical records. When calculating the degree of coincidence with the historical records, compare the large itemset with the historical records, and prune the records lower than the average value [29].

Algorithm 2 personalized pruning algorithm $prue(I_i, H_{user})$

```
Input: $I_i, H_{user}, I_x \in I_i$;
Output:
1. $num = \text{count}(I_i)$; $num$ is the number of large itemset items.
2. For $\omega = 1$: num.
3. $T_1 = total(I_w)$; $I_w = \{I_x^w\}, I_x^w \in I_x$.
4. $T_2 = count(H_{user})$; $T_2$ is the number recorded in $H_{user}$.
5. $T_3 = T_1/T_2$.
6. $T_4 = num/count(H_{user})$.
7. If $(T_3 < T_4)$; less than the average are pruned.
Delete $I_w$; Delete $I_x$ from $I_i$.
8. Return$(I_R)$.
End
```

T_4 in algorithm 2 is the threshold selected for this study [30].

Generally, the recommended results can be evaluated according to accuracy and re call [31]. Recall rate refers to the proportion of recommended books that meet readers' interests in the concentration of readers' interests; Accuracy refers to the proportion of recommended books in line with readers' interests in the total recommended book collection [32]. The calculation formulas of recall rate and accuracy rate are shown in formula 4, formula 5 and formula 6.

Recall rate: $$R_e = \sum_{i=1}^{u} \frac{L_i}{M \times P_i} \quad (4)$$

Accuracy: $$P_r = \frac{\sum_{i=1}^{M} L_i}{M \times N} \quad (5)$$

Harmonic average of the two: $$F = \frac{2 \times R_e \times P_r}{R_e + P_r} \quad (6)$$

Where, P_i is the total number of books in the reader's interest concentration, L_i is the recommended books that meet the reader's interest, M is the total number of readers, and N is the total number of recommended books. The higher the F value, the better the recommendation effect [33].

4 Results and analysis

This section presents the result analysis obtained from the proposed recommendation model and presents its discussion and summary in conclusion section.

The Bayesian algorithm is tested, and the accuracy, recall and F index are used to judge the quality of recommendation. The specific design of the experiment is as follows. In this experiment, the number of neighbors is 10. The change trend of F index, accuracy and recall is shown in Figures 3, 4 and 5 [35, 35].

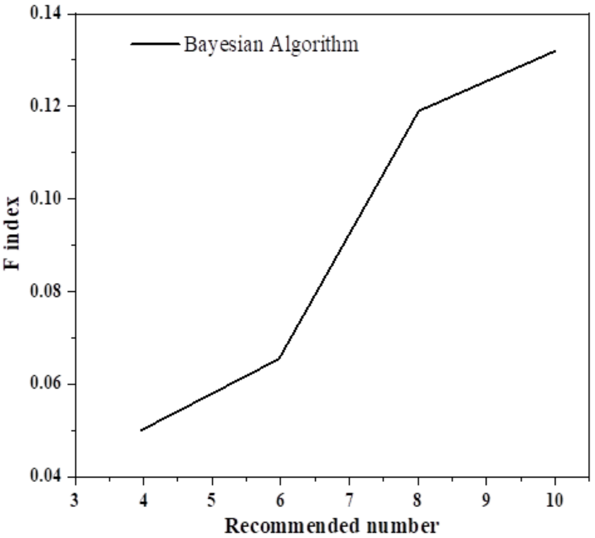


Figure 3: Change trend of F index.

It can be seen from Figure 3 that the f index based on Bayesian algorithm reaches about 14%, so it can be concluded that the recommendation effect based on Bayesian algorithm is better.

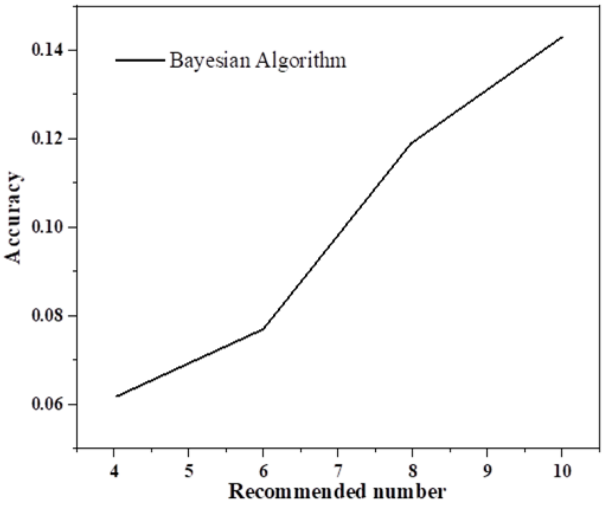


Figure 4: Variation Trend of accuracy.

As can be seen from Figure 4, the highest accuracy based on Bayesian algorithm is about 15%. As can be seen from Figure 5, the highest recall rate based on Bayesian algorithm is about 13%. The hybrid recommendation algorithm combining the recommendation results of Apriori algorithm and Bayesian algorithm is designed to judge the recommendation effect of the hybrid recommendation algorithm. The average absolute error, accuracy, recall and f index are used to judge the recommendation quality [36, 37]. The experimental design and result analysis are described below.

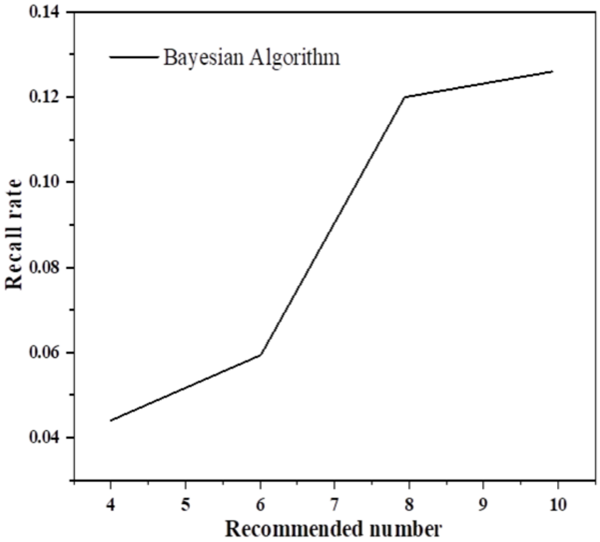


Figure 5: Change trend of recall rate.

Experiment 1: In this experiment, a hybrid recommendation algorithm combining Apriori algorithm and Bayesian algorithm is used to verify the recommendation effect of the hybrid recommendation algorithm. According to the design of Experiment 1, the change trend of F index, accuracy and recall of hybrid recommendation algorithm is shown in Figures 6, Figure 7 and Figure 8 [38, 39].

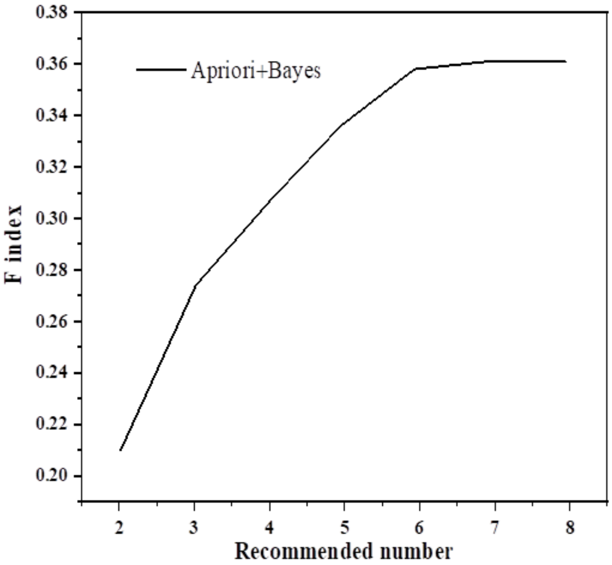


Figure 6: Change trend of F index.

As can be seen from Figure 6, the F index of the hybrid recommendation algorithm based on Apriori algorithm and Bayesian algorithm reaches about 34%, which is about 20% higher than that of the Bayesian algorithm alone, and improves the accuracy of recommendation to a certain extent.

As can be seen from Figure 7, the accuracy of the hybrid estimation method of Apriori algorithm and Bayesian algorithm is basically stable at about 33%, which is about 20% higher than that of Bayesian algorithm alone.

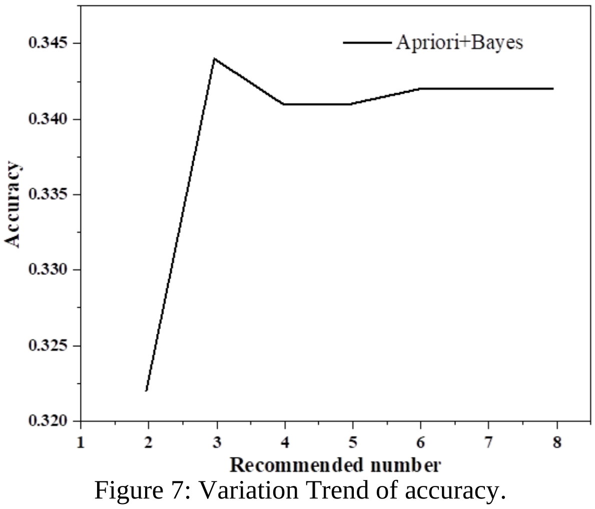


Figure 7: Variation Trend of accuracy.

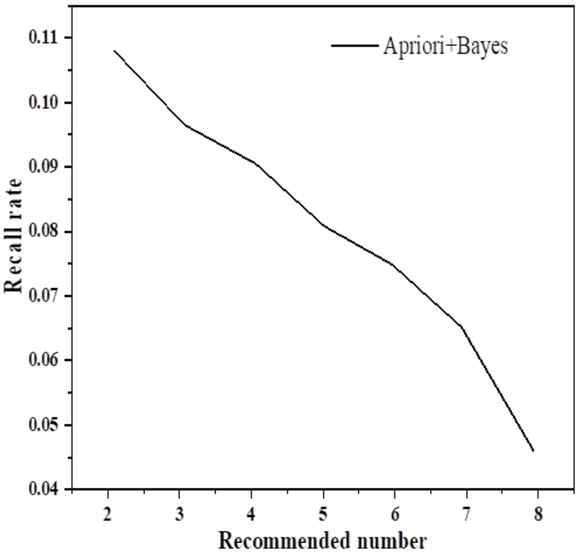


Figure 8: Change trend of recall rate.

As can be seen from figure 8, the highest recall rate of Apriori algorithm and Bayesian algorithm is about 11%, which is about 2% lower than that of Bayesian algorithm alone, which fully shows that better results can be obtained by combining the two algorithms.

5 Conclusions

This study successfully combines user history information with Bayesian network verification, and obtains an effective personalized Association recommendation model. The proposed model can solve the problem of weak personalization in the existing recommendation system to a certain extent. The model can eliminate the recommended goods with low probability of "preference" and highlight the goods with high probability of "preference", so as to give priority to the learning resources that readers really like. Further research includes optimizing the pruning method of association rule mining from user historical data, improving the mining algorithm of association rules. It describes the personalization and semantics, such as the method of introducing ontology. The research on personalized recommendation of e-learning resources is fundamental from the perspective of education, which is the necessity of the development of technology and learning environment. To fundamentally solve learners' learning needs and improve the effect of recommendation, it is inseparable from the guidance of educational theory. Secondly, the research on personalized recommendation of e-learning resources should be combined with the educational process. Learners' needs are not only for specific learning resources, but also for various personalized learning services in the learning process. The research on personalized recommendation of e-learning resources is still in the development stage. The research perspective for the future scope should focus on the application of these system models in specific education and teaching practice, so as to fundamentally promote teaching reform and innovation.

Acknowledgement:

Shanxi philosophy and Social Science Program "Research on the construction of accurate recommendation model of Digital Library Based on knowledge map" "(Project No.: 2020YJ148)".

References

[1] Cong, H. (2020). Personalized recommendation of film and television culture based on an intelligent classification algorithm. *Personal and Ubiquitous Computing*, *24*(2), 165-176. *https://doi.org/10.1007/s00779-019-01271-8*

[2] Shu, J., Shen, X., Liu, H., Yi, B., & Zhang, Z. (2018). A content-based recommendation algorithm for learning resources. *Multimedia Systems*, *24*(2), 163-173. *https://doi.org/10.1007/s00530-017-0539-8*

[3] Saito, T., & Watanobe, Y. (2020). Learning path recommendation system for programming education based on neural networks. *International Journal of Distance Education Technologies (IJDET)*, *18*(1), 36-64. *10.4018/IJDET.2020010103*

[4] Rawat, B., & Dwivedi, S. K. (2019). Discovering Learners' characteristics through cluster analysis for recommendation of courses in E-learning environment. *International Journal of Information and Communication Technology Education (IJICTE)*, *15*(1), 42-66. *10.4018/IJICTE.2019010104*

[5] Zhong, W. (2019). Design and Application of Scratch Personalized Learning Resources Based on VAK Learning Style Theory—Take Q School for Example. *Open Journal of Social Sciences*, *7*(8), 346-361. *10.4236/jss.2019.78025*

[6] Benhamdi, S., Babouri, A., & Chiky, R. (2017). Personalized recommender system for e-Learning environment. *Education and Information Technologies*, *22*(4), 1455-1477. *https://doi.org/10.1007/s10639-016-9504-y*

[7] Chao, L., Wen-hui, Z., & Ji-ming, L. (2019). Study of star/galaxy classification based on the xgboost algorithm. *Chinese Astronomy and Astrophysics*, *43*(4), 539-548. *https://doi.org/10.1016/j.chinastron.2019.11.005*

[8] Kumar, A., Pandey, D. S., & Namdeo, V. (2019). A survey on finding network traffic classification methods based on c5.0 machine learning algorithm. International journal of computer sciences and engineering, 7(4), 788-791. *10.26438/ijcse/v7i4.788791*

[9] Sharma, S., & Suryavanshi, A. An Efficient Personalized POI Recommendation using PCA-SVM based Filtering and Classification. *International Journal of Computer Applications*, *975*, 8887. *10.5120/ijca2017915801*

[10] Zhang, Y., Zhou, G., Jin, J., Zhao, Q., Wang, X., & Cichocki, A. (2015). Sparse Bayesian classification of EEG for brain–computer interface. *IEEE transactions on neural networks and learning systems*, *27*(11), 2256-2267. *10.1109/TNNLS.2015.2476656*

[11] Chinna Gounder Dhanajayan, R., & Appavu Pillai, S. (2017). SLMBC: spiral life cycle model-based Bayesian classification technique for efficient software fault prediction and classification. *Soft Computing*, *21*(2), 403-415. *https://doi.org/10.1007/s00500-016-2316-6*

[12] Zhu, Y., Li, X., Wang, J., Liu, Y., & Qu, Z. (2017). Practical secure naïve bayesian classification over encrypted big data in cloud. *International Journal of Foundations of Computer Science*, *28*(06), 683-703. *https://doi.org/10.1142/S0129054117400135*

[13] Zhang, Y., Zhou, G., Jin, J., Zhao, Q., Wang, X., & Cichocki, A. (2015). Sparse Bayesian classification of EEG for brain–computer interface. *IEEE transactions on neural networks and learning systems*, *27*(11), 2256-2267. *10.1109/TNNLS.2015.2476656*

[14] Kumbinarasaiah, S., & Raghunatha, K. R. (2021). A novel approach on micropolar fluid flow in a porous channel with high mass transfer via wavelet frames. *Nonlinear Engineering*, *10*(1), 39-45. *https://doi.org/10.1515/nleng-2021-0004*

[15] Dhotre, P. K., & Srinivasa, C. V. (2021). On free vibration of laminated skew sandwich plates: A finite element analysis. *Nonlinear Engineering*, *10*(1), 66-76. *https://doi.org/10.1515/nleng-2021-0006*

[16] Pedram, L., & Rostamy, D. (2021). Numerical simulations of stochastic conformable space–time fractional Kortewegde Vries and Benjamin–Bona–Mahony equations. *Nonlinear Engineering*, *10*(1), 77-90. *https://doi.org/10.1515/nleng-2021-0007*

[17] Wang, H., Hao, L., Sharma, A., & Kukkar, A. (2022). Automatic control of computer application data processing system based on artificial intelligence. *Journal of Intelligent Systems*, *31*(1), 177-192. *https://doi.org/10.1515/jisys-2022-0007*

[18] Sun, L., Gupta, R. K., & Sharma, A. (2022). Review and potential for artificial intelligence in healthcare. *International Journal of System Assurance Engineering and Management*, *13*(1), 54-62. *https://doi.org/10.1007/s13198-021-01221-9*

[19] Cai, Y., & Sharma, A. (2021). Swarm intelligence optimization: an exploration and application of machine learning technology. *Journal of Intelligent Systems*, *30*(1), 460-469. *https://doi.org/10.1515/jisys-2020-0084*

[20] Wang, Y., & Pedram, M. (2016). Model-free reinforcement learning and bayesian classification in system-level power management. *IEEE Transactions on Computers*, *65*(12), 3713-3726.

10.1109/TC.2016.2543219

[21] Liu, Y. (2020). Optimization of architectural art teaching model based on Naive Bayesian classification algorithm and fuzzy model. *Journal of Intelligent & Fuzzy Systems*, *39*(2), 1965-1976. *10.3233/JIFS-179966*

[22] Liang, Y., Xing, Y., & Zhang, Q. (2017). Parallel scheduling algorithm with improved bayesian classification algorithm. Wutan Huatan Jisuan Jishu, 39(3), 411-415

[23] Liang, N., Sun, S., Zhang, C., He, Y., & Qiu, Z. (2022). Advances in infrared spectroscopy combined with artificial neural network for the authentication and traceability of food. *Critical Reviews in Food Science and Nutrition*, *62*(11), 2963-2984. *https://doi.org/10.1080/10408398.2020.1862045*

[24] Tiancheng, L., Qing-dao-er-ji, R., & Ying, Q. (2019). Application of Improved Naive Bayesian-CNN Classification Algorithm in Sandstorm Prediction in Inner Mongolia. *Advances in Meteorology*, *2019*. *https://doi.org/10.1155/2019/5176576*

[25] Shen, Z., Zhang, Y., & Chen, W. (2019). A bayesian classification intrusion detection method based on the fusion of PCA and LDA. *Security and Communication Networks*, *2019*. *https://doi.org/10.1155/2019/6346708*

[26] Rastogi, N., Rastogi, S., & Darbari, M. (2019). A Novel Software Reliability Prediction Algorithm Using Fuzzy Attribute Clustering and Nave Bayesian Classification. *International Journal of Computer Sciences and Engineering*, *7*(2), 73-82. *10.26438/ijcse/v7i2.7382*

[27] Bountris, P., Topaka, E., Pouliakis, A., Haritou, M., Karakitsos, P., & Koutsouris, D. (2016). Development of a clinical decision support system using genetic algorithms and Bayesian classification for improving the personalised management of women attending a colposcopy room. *Healthcare technology letters*, *3*(2), 143-149. *https://doi.org/10.1049/htl.2015.0051*

[28] Balaram, A., & Vasundra, S. (2022). Prediction of software fault-prone classes using ensemble random forest with adaptive synthetic sampling algorithm. *Automated Software Engineering*, *29*(1), 1-21. *https://doi.org/10.1007/s10515-021-00311-z*

[29] Zherdev, I. Y. (2020). Assessment of Bayesian Ternary Gaze Classification Algorithm (I-BDT). *Modelling and Data Analysis*, *10*(2), 74-92. *https://orcid.org/0000-0001-6810-9297*

[30] Nykaza, E. T., Blevins, M. G., Hart, C. R., & Netchaev, A. (2017). Bayesian classification of environmental noise sources. *The Journal of the Acoustical Society of America*, *141*(5), 3522-3522. *https://doi.org/10.1121/1.4987416*

[31] Karabatsos, G. (2021). Fast Search and Estimation of Bayesian Nonparametric Mixture Models Using a Classification Annealing EM Algorithm. *Journal of Computational and Graphical Statistics*, *30*(1), 236-247. *https://doi.org/10.1080/10618600.2020.1807995*

[32] Mittal, A. K., Mittal, S., & Rathore, D. S. (2018). Bayesian Classification for Social Media Text. International journal of computer sciences and engineering, 6(7), 641-646.

[33] Zhu, W. (2020). Classification accuracy of basketball simulation training system based on sensor fusion and Bayesian algorithm. *Journal of Intelligent & Fuzzy Systems*, *39*(4), 5965-5976. *10.3233/JIFS-189070*

[34] Xu, Y., Zhang, J., Gong, X., Jiang, K., Zhou, H., & Yin, J. (2016). A method of real-time traffic classification in secure access of the power enterprise based on improved random forest algorithm. *Power System Protection and Control*, *44*(24), 82-89. *10.7667/PSPC152144*

[35] Sharma, R., Raju, C. S., Animasaun, I. L., Santhosh, H. B., & Mishra, M. K. (2021). Insight into the significance of Joule dissipation, thermal jump and partial slip: dynamics of unsteady ethelene glycol conveying graphene nanoparticles through porous medium. *Nonlinear Engineering*, *10*(1), 16-27. *https://doi.org/10.1515/nleng-2021-0002*

[36] Xingrong, S. (2019). Research on time series data mining algorithm based on Bayesian node incremental decision tree. *Cluster Computing*, *22*(4), 10361-10370. *https://doi.org/10.1007/s10586-017-1358-6*

[37] Chen, Y., Zhang, W., Dong, L., Cengiz, K., & Sharma, A. (2021). Study on vibration and noise influence for optimization of garden mower. *Nonlinear Engineering*, *10*(1), 428-435. *https://doi.org/10.1515/nleng-2021-0034*

[38] Darwish, S. M. (2016). Combining firefly algorithm and Bayesian classifier: new direction for automatic multilabel image annotation. *IET Image Processing*, *10*(10), 763-772. *https://doi.org/10.1049/iet-ipr.2015.0492*

[39] Sharma, A., Georgi, M., Tregubenko, M., Tselykh, A., & Tselykh, A. (2022). Enabling smart agriculture by implementing artificial intelligence and embedded sensing. *Computers & Industrial Engineering*, *165*, 107936. *https://doi.org/10.1016/j.cie.2022.107936*

Data Processing of Municipal Wastewater Recycling Based on Genetic Algorithm

Zijun Zhao[1,2], Jianchao Zhu[2], Kaiming Yang[1*], Song Wang[1,2], Mingxiao Zeng[1,2]
[1]College of Civil Architecture and Environment, Xihua University, Chengdu 610039, China
[2]Chinese Research Academy of Environmental Sciences, Beijing,100012, China
E-mail: zijunzhao3@126.com, jianchaozhu@163.com, kaimingyang3@126.com, songwang839@163.com, mingxiaozeng6@126.com

Keywords: genetic algorithm, adaptive algorithm, PID control, urban sewage, sewage regeneration and reuse, dissolved oxygen control

Received: February 24, 2022

This paper designs an adaptive genetic algorithm in order to accurately process the data of urban sewage recycling. The proposed algorithm integrates genetic algorithm, adaptive genetic algorithm, traditional PID respectively, and designs simulation experiments to compare their performance. The simulation results show that the self-adaptive PID control algorithm is superior to the genetic PID control algorithm in both control accuracy and dynamic characteristics. The PID controller with good optimization performance is applied to the control object of sewage treatment system. Through simulation analysis, the adaptive genetic algorithm only needs 52s when adjusting the step response simulation. The overshoot of the system is observed as 8% which is better in comparison with existing baseline model. The interference in the simulation is restored to a stable state within the interference 18s, and the adjustment time in the robustness simulation is reduced by about 15s compared with the genetic algorithm. In conclusion, the adjustment time of the system is shortened, the overshoot of the system is reduced, and the anti-interference and robustness are enhanced. For the dissolved oxygen concentration of the key object in the control system, the above controller with good performance is applied to the sewage treatment control system, which not only reduces the overshoot and regulation time, but also improves the control accuracy, and can well meet the control requirements of sewage treatment.

Povzetek: Članek predstavlja prilagodljiv genetski algoritem za obdelavo podatkov mestne kanalizacije, ki izboljšuje natančnost in robustnost pri obdelavi odpadnih vod ter optimizira regulacijo kisika.

1 Introduction

With the continuous growth of China's population and the rapid development of economy, the water consumption and drainage are increasing year by year, and the limited water resources are continuously polluted. In addition, the uneven distribution of regional water resources and periodic drought lead to the increasingly acute contradiction between supply and demand of water resources. The shortage of water resources has become the bottleneck restricting China's social and economic development. For a long time, people used to discharge the once used water directly. It's incredible that it has other uses. In fact, water is the only irreplaceable resource in nature, and it is also a renewable resource. Of the water used by people, only about 0.1% is polluted by impurities (compared with 3.5% in seawater), and most of the rest can be reused. After proper regeneration treatment, sewage can be reused to realize a virtuous cycle of water in nature. Urban sewage is available nearby, easy to collect and treat, has a huge quantity and stable and reliable source, is not affected by natural factors such as climate, and there is no dispute over the right to anhydrous resources. As the second water source of the city, sewage treatment and recycling is more economical than long-distance water diversion or water transfer, seawater desalination and so on. The extensive utilization of urban reclaimed water can not only reduce the water intake to the natural water body, but also reduce the pollution load discharged to the natural water body. The purpose of exploring the optimization of urban reclaimed water system is to promote the scientific and reasonable planning, construction and operation of urban reclaimed water system [1-3].

The urban reclaimed water system consists of block sewage pipe network, municipal sewage pipe network, sewage lifting pump station, reclaimed water plant, reclaimed water booster pump station, municipal reclaimed water pipe network and block reclaimed water pipe network. Figure 1 shows the composition of urban reclaimed water system.

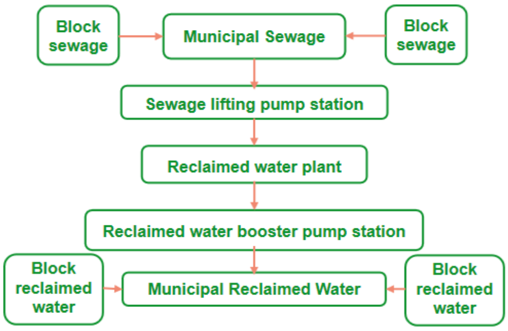


Figure 1: Composition diagram of urban reclaimed water system.

Genetic algorithm has a good effect on parameter optimization. It is an algorithm that imitates the evolution of natural organisms. Because genetic algorithm has good parameter optimization effect, it has been widely studied and applied in PID control system [4-5]. However, the application of genetic algorithm in PID control system has a series of shortcomings, such as easy precocity and slow convergence speed. Aiming at the above problems of genetic algorithm, this paper designs an adaptive genetic algorithm, which can retain excellent individuals and automatically adjust the crossover and mutation probability according to individual conditions.

In this paper, genetic algorithm, adaptive genetic algorithm and traditional PID are fused together, and simulation experiments are designed to compare their performance. The simulation results show that the self-adaptive PID control algorithm is superior to the genetic PID control algorithm in both control accuracy and dynamic characteristics. The PID controller with good optimization performance is applied to the control object of sewage treatment system. The rest of this article is systematized as literature is presented in section 2 followed by research methods in section 3. Section 4 depicts the results and the conclusion is presented in section 5.

2 Related work

In this section various state-of-the-art work in the field of wastewater treatment using several approaches are discussed.

As the main secondary sewage biochemical treatment technology, activated sludge process is widely used all over the world because of its strong anti-interference ability, wide treatment range, fast treatment speed and relatively low cost. Activated sludge includes microorganisms in water and substances attached to microbial communities. Activated sludge treatment is composed of two process parts: biological aeration tank treatment process part and secondary sedimentation tank treatment process part. Through the metabolism of bacteria and other microorganisms in activated sludge, it centrally adsorbs and oxidizes and decomposes the polluting organic substances in sewage, so as to purify water quality [6]. The overall occurrence process is shown in Figure 2 below.

In recent years, PID controller based on genetic algorithm optimization has become a research hotspot of scholars at home and abroad. Excellent research materials introducing the application of genetic algorithm to PID parameter tuning emerge one after another. Genetic PID algorithm has been widely used in theoretical basis and engineering research, and has achieved a lot of results. These research results have proved that compared with the traditional PID tuning, the optimization tuning based on genetic algorithm has better practicability and optimization [7]. Li *et al.* [8] proposed a hybrid genetic algorithm based on bacterial foraging algorithm. When adjusting the PID control parameters of AVR, this algorithm is used. The key research is on the variation trend of variation, crossover, mutation step size and crossover step size. Then, the results of simulation experiments show that the algorithm has good anti-interference performance.

Hernandez *et al.* [9] integrates genetic algorithm and PID control algorithm for parameter optimization design, applies this intelligent algorithm to distributed parameter objects, uses DP method when calculating the parameter stability region of the control system, and compares it with several control methods using conventional parameter setting formula to obtain the comparison results. MATLAB software is used in the simulation experiment. The simulation results show that the algorithm is effective and feasible. Hu *et al.* [10] proposed a PID control algorithm based on quantum genetic algorithm. In order to achieve the purpose of population evolution, this method uses the individual representation of quantum bits and quantum revolving gate, which can realize PID multi-objective optimization, and proves the feasibility of parameter tuning. Ao *et al.* [11] studied the single neuron control algorithm based on genetic algorithm, which improved the calculation efficiency and convergence speed, gradually reduced the search space and found the best data when the population number and crossover probability were decreasing. The simulation results show that the single neuron control algorithm based on genetic algorithm has good parameter optimization effect.

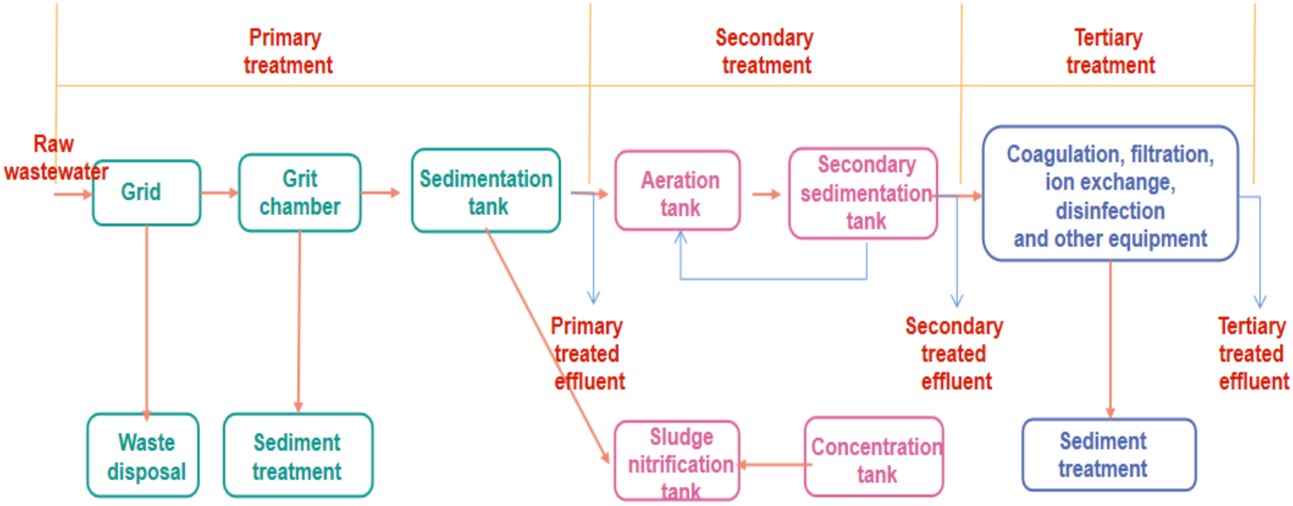


Figure 2: Typical process flow of activated sludge.

3 Research methods

This section includes the optimization process and simulation process of proposed genetic algorithm for wastewater treatment system.

3.1 Genetic algorithm parameter optimization

Genetic algorithm is a process of repeatedly searching for more optimized regions. Through this “guidance”, excellent new individuals will continue to emerge, and inferior individuals will be quickly eliminated. Thus, organisms evolve to a higher stage [12]. Figure 3 is the basic flow chart of genetic algorithm.

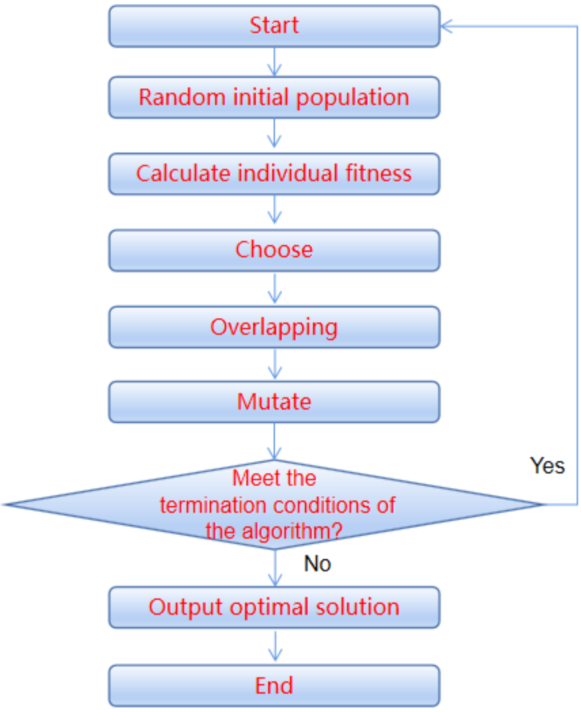

Start
Random initial population
Calculate individual fitness
Choose
Overlapping
Mutate
Meet the termination conditions of the algorithm?
Yes
No
Output optimal solution
End

Figure 3: Flow chart of genetic algorithm.

The shortcomings of genetic algorithm are mainly reflected in the following points:

i. Due to the lack of local search and fine-tuning ability of genetic algorithm, it is difficult to determine the exact position of the optimal solution, which makes the genetic algorithm converge too early and cannot achieve the purpose of optimization.

ii. Some genes of the original old population may change due to random variation, which will affect the speed of the algorithm converging to the excellent solution to a certain extent. Because the generation of offspring individuals is random and the crossover operation is the same for all individuals, it is not guaranteed that the offspring will be better than the parent individuals.

iii. The pattern diversity of genetic algorithm is difficult to maintain, which is easy to lead to premature convergence, so that the optimization effect is not very ideal [13].

The above problems often appear in the practical application of genetic algorithm. In order to improve the convergence and convergence speed of genetic algorithm, some improvement measures are made on Genetic Algorithm in theoretical research and practical application. In order to improve genetic algorithm, it is usually necessary to determine the coding parameter scheme, set the appropriate population size, genetic algorithm structure, and select the appropriate Pc and Pm.

When operating genetic algorithm, crossover and mutation are the main factors that determine the convergence performance of the algorithm. Crossover and mutation directly affect and determine the convergence of the algorithm. An improved adaptive crossover and mutation operator is proposed in this paper. The basic idea of adaptive genetic algorithm is that when the genetic algebra is increasing, the operation mode of mutation operator and crossover operator will be automatically adjusted on the basis of genetic algorithm.

This paper mainly improves the adaptive genetic algorithm from the two aspects of adaptive mutation, the design of crossover probability and how to retain excellent individuals.

The probability value needs to be obtained through repeated experiments, the parameter optimization process is cumbersome, the efficiency is very low, and the optimal solution cannot meet all the conditions. Adaptive genetic algorithm can adjust PC and PM according to the fitness value during operation. Equation 1 is the dynamic adjustment of parameter crossover operator probability Pc of adaptive genetic algorithm, and Equation 2 is the dynamic adjustment of parameter mutation operator probability Pm of adaptive genetic algorithm.

$$P_C = \begin{cases} p_{c1} - \dfrac{(p_{c1} - p_{c2})(F' - F_{max})}{F_{max} - F_{avg}}, F' \geq F_{max} \\ p_{c1}, F' < F_{max} \end{cases} \quad (1)$$

$$P_m = \begin{cases} p_{m1} - \dfrac{(p_{m1} - p_{m2})(F' - F_{max})}{F_{max} - F_{avg}}, F' \geq F_{max} \\ p_{m1}, F' < F_{max} \end{cases} \quad (2)$$

In the above two Equations 1 and 2: Fmax represents the maximum value of fitness function in each generation of individuals. Favg represents the average value of fitness function in each generation; F 'represents the larger of the fitness function values of the two paired individuals. Equations 1 and 2 show that when the fitness values of most individuals in the population are concentrated, PC and PM are large, and the adaptability of individuals whose fitness is lower than the average fitness of the population is poor; When the distribution range of individual fitness value in the population is large, Pc and Pm are small, and the individual whose fitness is higher than the average fitness of the population has better adaptability. For individuals whose fitness is almost the same as the average fitness of the population, their Pc and Pm are almost equal to 0. Adaptive genetic algorithm can increase individual fitness, improve the overall quality of the population, enhance the diversity of the population, and improve the ability of searching close to the optimal solution.

3.2 Application and Simulation of genetic algorithm PID control in sewage treatment system

DO (dissolved oxygen) refers to the oxygen combined with water in molecular form, which can directly affect the water quality. During the biochemical treatment of sewage, the compressed air is sent to the aeration head by the blower through the air supply pipe. The aeration head continuously turns the air into micro bubbles and enters the water, resulting in violent mixing and stirring of water in the tank, increasing the contact surface of sludge, making the sewage fully contact with microbiota, and promoting the combination of oxygen and water to form dissolved oxygen. Dissolved oxygen raises enough oxygen for cells. Temperature, air pressure and salt content in water will affect the content of dissolved oxygen. Dissolved oxygen mainly provides oxygen for the oxidation and decomposition of organic matter in sewage, and some reducing substances also need some oxygen. Therefore, dissolved oxygen is particularly important for sewage treatment. Dissolved oxygen with a concentration of about 2mg / L is often used to ensure the effective removal of organic matter and self-survival of microbial bacteria [14]. Dissolved oxygen parameters need to be studied in this system. In the process of dissolved oxygen control, the conventional PID control cannot adjust the control parameters well and cannot adapt to the system changes, resulting in poor regulation effect. If genetic algorithm is combined with the former, the above problems can be overcome. In order to better adapt to the changing parameters and working conditions, genetic algorithm and PID control can be combined to improve the whole control. Therefore, the above methods are adopted to control the dissolved oxygen (DO) concentration in the sewage treatment system to ensure that the sewage can be treated to meet the standard with the lowest energy consumption. When the control system is running, the blower sends the air from the outside to the biochemical tank through the pipeline to deliver oxygen to the tank to improve the concentration of dissolved oxygen. When adjusting the dissolved oxygen concentration, it is only necessary to adjust the speed of the blower, control the wind speed and control the air supply volume, so as to achieve the purpose of dissolved oxygen control. Therefore, the control of dissolved oxygen can be transformed into the control of blower. The structure diagram of dissolved oxygen control system is depicted in Figure 4.

As shown in Figure 4, the whole dissolved oxygen control system consists of three parts: aeration flow control link (composed of frequency converter and blower), aeration mass transfer process and dissolved oxygen detection link. DOa, Doset, DOc and Q are the actual value of dissolved oxygen, the set value of dissolved oxygen, the measured value of dissolved oxygen and the flow of air blown by the blower [15]. The system calculates the difference between the detected value of dissolved oxygen and the set value of dissolved oxygen, and uses the PID controller to calculate the difference between the two. The PID controller adjusts the frequency of the output control quantity of the frequency converter controlling the blower speed, so as to realize the control of the air supply volume of the blower, and then achieve the control of the dissolved oxygen concentration [16].

According to the material balance formula as shown in Equation 3, DO change rate = DO input rate - DO output rate - DO consumption rate, the following activated sludge dynamic model can be established [17].

$$V\frac{dc}{dt} = QC_0 - QC_1 - VkC \quad (3)$$

$$G(s) = \frac{Q(s)}{C(s)} = \frac{C_0 - C_1}{VS + VK} = \frac{(C_0 - C_1)/V}{S + K} \tag{4}$$

Let(C0-C1)/V = R, then Equation (4) becomes G(S) = W/(S+K), which is an inertial link.

The commonly used measurement method of dissolved oxygen concentration is diaphragm electrode method. According to the measurement principle of do instrument, i.e., electrochemical equation.

$$DO_{nt} = DO_{st} + \frac{\gamma}{a}(1 - e^{-at}) \tag{5}$$

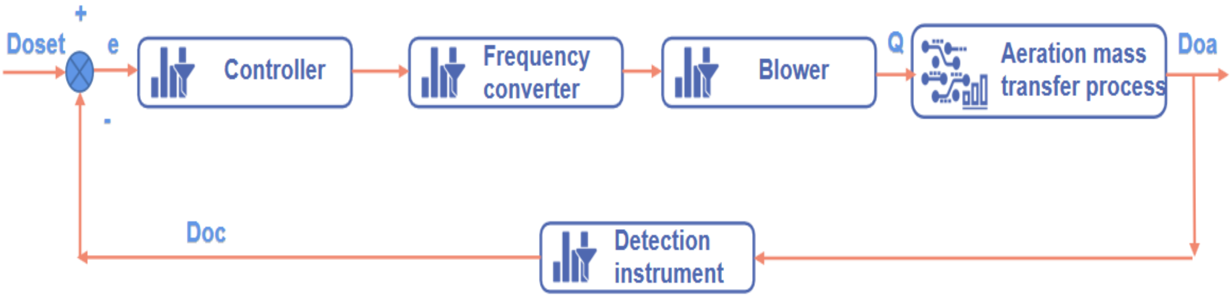


Figure 4: Structure of dissolved oxygen control system.

In Equation 5, DO_{nt} is the measured value of DO at time t; Dost is the actual DO concentration in the water sample at time t; γ is the rate of microbial oxygen consumption; A is the parameter of electrode response speed of do instrument, min-1.

At high concentration, a = 9.2 min-1. It can be seen that the detection of do is nonlinear and has lag characteristics. The model is modified by Equation 3, and the detection lag is expressed by pure lag τ, then Equation 4 is modified as Equation 6.

$$G(s) = \frac{R}{S + K} e^{-\tau s} \tag{6}$$

According to the modified model, the dissolved oxygen treatment process is approximately a first-order inertial pure lag link [18].

According to the mathematical model of dissolved oxygen concentration and experience, the dissolved oxygen control model [19] is selected as Equation 7.

$$G(s) = \frac{15.24}{S + 0.0157} e^{-15s} \tag{7}$$

In order to compare the application effects of the three control algorithms, the three algorithms are simulated and analyzed in this paper.

The initial parameter value obtained by Z-N setting method is:

Kp=0.742, Ki=0.001, Kd=2.123;

Sample size is 30;
Evolutionary algebra is 100;
Probability of crossover operator is Pc = 0.9;
Probability of mutation operator is Pm = 0.03.

The initial parameter value of PID is used to set the parameter range, and the parameters are optimized in the parameter range. The PID controller based on genetic algorithm and the PID controller based on adaptive genetic algorithm are used to set and optimize the parameters respectively. The control characteristics and optimization effects of these optimization methods are compared through simulation analysis.

4 Results and analysis

This section describes the step response simulation, anti-interference simulation and robust simulation for measuring the performance of proposed system.

4.1 Step response simulation

The step response curve simulated by MATLAB is shown in Figure 5 below.

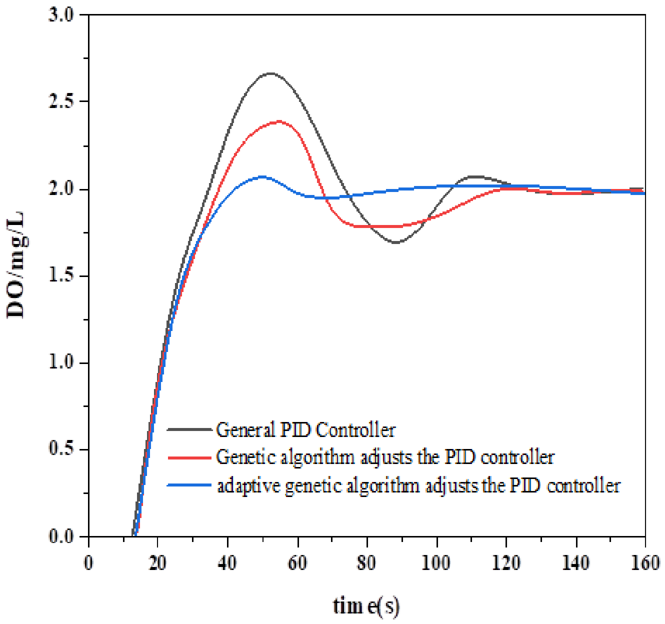


Figure 5: Step simulation comparison curve.

In terms of the rise time of the system, curve 1 takes 20s, the longest time, curve 2 takes 18S, while curve 3 takes the shortest time, only 15s, which can be increased from 0.2mg/l to 1.8mg/l. Curve 3 rises the most slowly and stably, curve 1 rises the most, and curve 2 takes the second place.

The adjustment time of curve 3 only needs 52s, the overshoot of the system is 8%, and the effect is the best, while the adjustment time of curve 2 and curve 1 are 118.8s and 103.5s respectively, and the corresponding overshoot is 28% and 20% respectively. At 52s, curve 1 and curve 2 begin to decline after reaching the maximum value and are in a fluctuating state. It will reach equilibrium at about 120s. At this time, curve 3 has reached equilibrium and has the best effect in optimization [20].

4.2 Anti-interference simulation

At 115 seconds, add -0.45mg/l interference to the output of the system. The simulation results are shown in Figure 6 below.

When the disturbance signal is added to the output of the system, the disturbance signal makes the simulation curves of the three control methods offset, and the offset of curve 1 is the largest. According to the PID parameter optimization disturbance characteristic curve 2 based on genetic algorithm, the system can recover to a stable state within 30s after being disturbed, and according to the PID parameter optimization disturbance characteristic curve 3 based on adaptive genetic algorithm, the system can recover to a stable state within 18S after being disturbed. By comparing the curves, it can be seen that the adaptive genetic algorithm represented by curve 3 has the best anti-interference performance and the shortest time for the system to recover to the stable state [21].

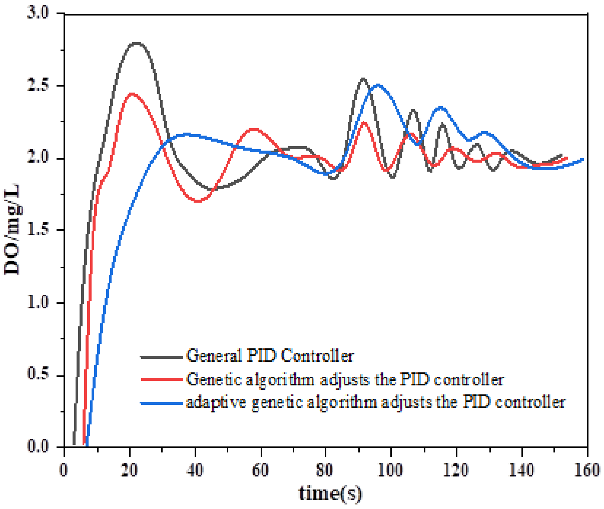


Figure 6: Comparison curve of anti-interference simulation.

4.3 Robustness simulation

When the dissolved oxygen tester works, it is immersed in sewage for a long time, which is easy to be corroded, resulting in mechanical passivation, change of test accuracy, slightly lengthen the measurement lag time, and even change of system parameters. Therefore, in order to ensure that the system can maintain a stable and efficient operation state for a long time, a robust control method should be selected.

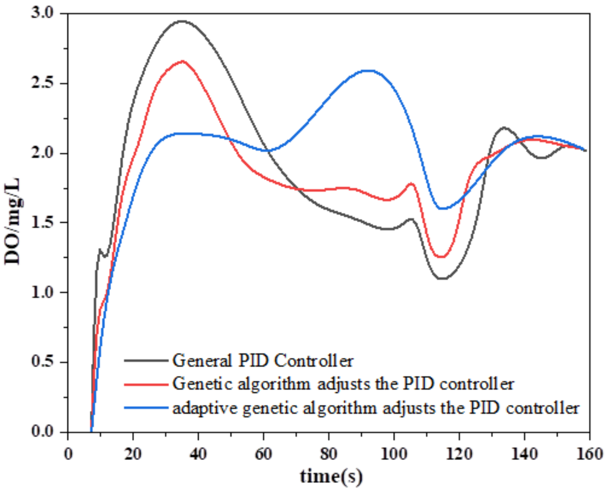


Figure 7: Comparison curve of robustness simulation.

Assuming that the dissolved oxygen aeration flow control link does not change, the steady-state gain of the aeration mass transfer link of the system increases by 75%, and its robustness is analyzed by the simulation curve 7.

The simulation curves of the three control methods have a great oscillation, and the oscillation amplitude of curve 1 is the largest. According to the comparison between curve 2 and curve 3, the adjustment time of the latter is reduced by about 15s compared with the former. Therefore, the adaptive genetic algorithm has better robustness and remarkable optimization effect.

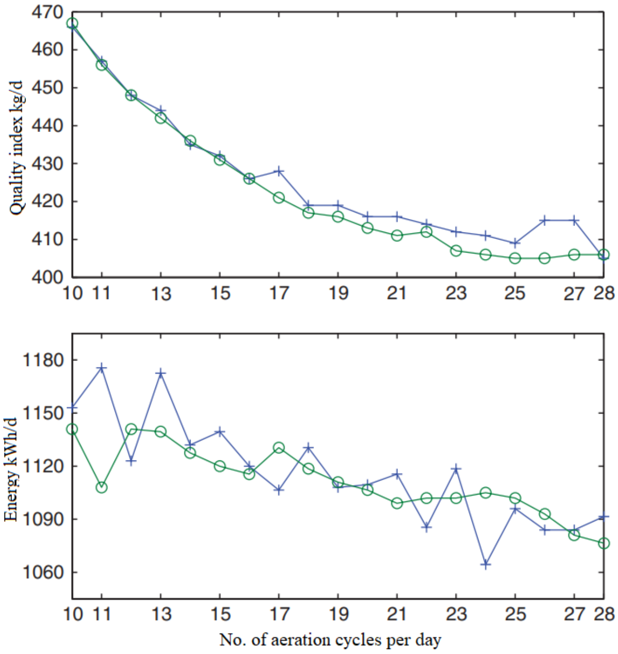


Figure 8: Optimization outcomes with respect to cycle numbers.

Ideal air circulation profiles have been figured out utilizing the recently portrayed optimization approach for 2NC factors applying simple identification. The relating enhanced values (EQ index and air circulation energy) can be seen in Figure 8. It very well may be seen that the EQ index is diminishing as the quantity of cycles increments. It was observed from the analysis that utilizing GA approach ideal arrangements can be proficiently found, moreover, the optimized outcome can diminish the contamination load with 10%.

5 Conclusions

This paper improves the problem of genetic algorithm and designs an adaptive genetic algorithm, which can retain excellent individuals and automatically adjust the crossover and mutation probability according to individual conditions. Genetic algorithm, adaptive genetic algorithm and traditional PID are fused together, and simulation experiments are designed to compare their performance. The simulation results show that the self-adaptive PID control algorithm is superior to the genetic PID control algorithm in both control accuracy and dynamic characteristics. The PID controller with good optimization performance is applied to the control object of sewage treatment system. The results show that the adjustment time of the system is shortened, the overshoot of the system is reduced, and the anti-interference and robustness are enhanced. The adaptive genetic algorithm only needs 52s when adjusting the step response simulation. The overshoot of the system is 8%. Interference simulation can be restored to stable state within 18S. In the robustness simulation, the adjustment time is reduced by about 15s compared with the genetic algorithm. For the dissolved oxygen concentration of the key object in the control system, the above controller with good performance is applied to the sewage treatment control system, which not only reduces the overshoot and regulation time, but also improves the control accuracy, and can well meet the control requirements of sewage treatment.

References

[1] Hamdi, H., Hechmi, S., Khelil, M. N., Zoghlami, I. R., Benzarti, S., Mokni-Tlili, S., & Jedidi, N. (2019). Repetitive land application of urban sewage sludge: Effect of amendment rates and soil texture on fertility and degradation parameters. Catena, 172, 11-20. https://doi.org/10.1016/j.catena.2018.08.015

[2] Gutiérrez-Alfaro, S., Rueda-Márquez, J. J., Perales, J. A., & Manzano, M. A. (2018). Combining sun-based technologies (microalgae and solar disinfection) for urban wastewater regeneration. Science of the Total Environment, 619, 1049-1057. https://doi.org/10.1016/j.scitotenv.2017.11.110

[3] Sabater-Liesa, L., Montemurro, N., Font, C., Ginebreda, A., González-Trujillo, J. D., Mingorance, N., & Barceló, D. (2019). The response patterns of stream biofilms to urban sewage change with exposure time and dilution. Science of The Total Environment, 674, 401-411. https://doi.org/10.1016/j.scitotenv.2019.04.178

[4] Nieuwenhuijse, D. F., Oude Munnink, B. B., Phan, M. V., Munk, P., Venkatakrishnan, S., Aarestrup, F. M., & Koopmans, M. P. (2020). Setting a baseline for global urban virome surveillance in sewage. Scientific Reports, 10(1), 1-13. https://doi.org/10.1038/s41598-020-69869-0

[5] Yu, Y. X., & Ahn, K. K. (2020). Energy regeneration and reuse of excavator swing system with hydraulic accumulator. International Journal of Precision Engineering and Manufacturing-Green Technology, 7(4), 859-873. https://doi.org/10.1007/s40684-019-00157-7

[6] Ifthikar, J., Jiao, X., Ngambia, A., Wang, T., Khan, A., Jawad, A., & Chen, Z. (2018). Facile one-pot synthesis of sustainable carboxymethyl chitosan–sewage sludge biochar for effective heavy metal chelation and regeneration. Bioresource technology, 262, 22-31. https://doi.org/10.1016/j.biortech.2018.04.053

[7] Zhou, Y., Zhang, Y., He, W., Wang, J., Peng, F., Huang, L., & Deng, W. (2018). Rapid regeneration and reuse of silica columns from PCR purification and gel extraction kits. Scientific reports, 8(1), 1-11. https://doi.org/10.1038/s41598-018-30316-w

[8] Li, X., Bardos, P., Cundy, A. B., Harder, M. K., Doick, K. J., Norrman, J., & Chen, W. (2019). Using a conceptual site model for assessing the sustainability of brownfield regeneration for a soft reuse: A case study of Port Sunlight River Park (UK). Science of The Total Environment, 652, 810-821. https://doi.org/10.1016/j.scitotenv.2018.10.278

[9] Hernández, L., Augusto, P. A., Castelo-Grande, T., & Barbosa, D. (2021). Regeneration and reuse of magnetic particles for contaminant degradation in water. Journal of Environmental Management, 285, 112155. https://doi.org/10.1016/j.jenvman.2021.112155

[10] Hu, Y., Zhao, C., Yin, L., Wen, T., Yang, Y., Ai, Y., & Wang, X. (2018). Combining batch technique with theoretical calculation studies to analyze the highly efficient enrichment of U (VI) and Eu (III) on magnetic MnFe2O4 nanocubes. Chemical Engineering Journal, 349, 347-357. https://doi.org/10.1016/j.cej.2018.05.070

[11] Ao, W., Fu, J., Mao, X., Kang, Q., Ran, C., Liu, Y., & Dai, J. (2018). Microwave assisted preparation of activated carbon from biomass: A review. Renewable and Sustainable Energy Reviews, 92, 958-979. https://doi.org/10.1016/j.rser.2018.04.051

[12] Palmieri, S., Cipolletta, G., Pastore, C., Giosuè, C., Akyol, Ç., Eusebi, A. L., & Fatone, F. (2019). Pilot scale cellulose recovery from sewage sludge and reuse in building and construction material. Waste Management, 100, 208-218. https://doi.org/10.1016/j.wasman.2019.09.015

[13] Akharame, M. O., Fatoki, O. S., & Opeolu, B. O. (2019). Regeneration and reuse of polymeric nanocomposites in wastewater remediation: the future of economic water management. Polymer Bulletin, 76(2), 647-681. https://doi.org/10.1007/s00289-018-2403-1

[14] Ye, T., Wang, K., Shuang, C., Zhang, G., & Li, A. (2019). Reuse of spent resin for aqueous nitrate removal through bio-regeneration. Journal of Cleaner Production, 224, 566-572. https://doi.org/10.1016/j.jclepro.2019.03.217

[15] Hermassi, M., Dosta, J., Valderrama, C., Licon, E., Moreno, N., Querol, X., & Cortina, J. L. (2018). Simultaneous ammonium and phosphate recovery and stabilization from urban sewage sludge anaerobic digestates using reactive sorbents. Science of the total environment, 630, 781-789. https://doi.org/10.1016/j.scitotenv.2018.02.243

[16] Li, J., Li, B., Huang, H., Zhao, N., Zhang, M., & Cao, L. (2020). Investigation into lanthanum-coated biochar obtained from urban dewatered sewage sludge for enhanced phosphate adsorption. Science of the Total Environment, 714, 136839. https://doi.org/10.1016/j.scitotenv.2020.136839

[17] Devane, M. L., Moriarty, E. M., Robson, B., Lin, S., Wood, D., Webster-Brown, J., & Gilpin, B. J. (2019). Relationships between chemical and microbial faecal source tracking markers in urban river water and sediments during and post-discharge of human sewage. Science of the Total Environment, 651, 1588-1604. https://doi.org/10.1016/j.scitotenv.2018.09.258

[18] Cabral, A. C., Wilhelm, M. M., Figueira, R. C., & Martins, C. C. (2019). Tracking the historical sewage input in South American subtropical estuarine systems based on faecal sterols and bulk organic matter stable isotopes (δ13C and δ15N). Science of The Total Environment, 655, 855-864. https://doi.org/10.1016/j.scitotenv.2018.11.150

[19] Bougnom, B. P., McNally, A., Etoa, F. X., & Piddock, L. J. (2019). Antibiotic resistance genes are abundant and diverse in raw sewage used for urban agriculture in Africa and associated with urban population density. Environmental Pollution, 251, 146-154. https://doi.org/10.1016/j.envpol.2019.04.056

[20] Gil-Meseguer, E., Bernabé-Crespo, M. B., & Gómez-Espín, J. M. (2019). Recycled sewage-a water resource for dry regions of Southeastern Spain. Water Resources Management, 33(2), 725-737. https://doi.org/10.1007/s11269-018-2136-9

[21] Gutiérrez-Alfaro, S., Rueda-Márquez, J. J., Perales, J. A., & Manzano, M. A. (2018). Combining sun-based technologies (microalgae and solar disinfection) for urban wastewater regeneration. Science of the Total Environment, 619, 1049-1057. https://doi.org/10.1016/j.scitotenv.2017.11.110

JOŽEF STEFAN INSTITUTE

Jožef Stefan (1835-1893) was one of the most prominent physicists of the 19th century. Born to Slovene parents, he obtained his Ph.D. at Vienna University, where he was later Director of the Physics Institute, Vice-President of the Vienna Academy of Sciences and a member of several sci- entific institutions in Europe. Stefan explored many areas in hydrodynamics, optics, acoustics, electricity, magnetism and the kinetic theory of gases. Among other things, he originated the law that the total radiation from a black body is proportional to the 4th power of its absolute tem- perature, known as the Stefan–Boltzmann law.

The Jožef Stefan Institute (JSI) is the leading independent scientific research institution in Slovenia, covering a broad spectrum of fundamental and applied research in the fields of physics, chemistry and biochemistry, electronics and information science, nuclear science technology, en- ergy research and environmental science.

The Jožef Stefan Institute (JSI) is a research organisation for pure and applied research in the natural sciences and technology. Both are closely interconnected in research de- partments composed of different task teams. Emphasis in basic research is given to the development and education of young scientists, while applied research and development serve for the transfer of advanced knowledge, contributing to the development of the national economy and society in general.

At present the Institute, with a total of about 900 staff, has 700 researchers, about 250 of whom are postgraduates, around 500 of whom have doctorates (Ph.D.), and around 200 of whom have permanent professorships or temporary teaching assignments at the Universities.

In view of its activities and status, the JSI plays the role of a national institute, complementing the role of the uni- versities and bridging the gap between basic science and applications.

Research at the JSI includes the following major fields: physics; chemistry; electronics, informatics and computer sciences; biochemistry; ecology; reactor technology; ap- plied mathematics. Most of the activities are more or less closely connected to information sciences, in particu- lar computer sciences, artificial intelligence, language and speech technologies, computer-aided design, computer architectures, biocybernetics and robotics, computer automa- tion and control, professional electronics, digital communications and networks, and applied mathematics.

The Institute is located in Ljubljana, the capital of the in dependent state of **Slove**nia (or S♡nia). The capital today isconsidered a crossroad bet between East, West and Mediter-ranean Europe, offering excellent productive capabilities and solid business opportunities, with strong international connections. Ljubljana is connected to important centers such as Prague, Budapest, Vienna, Zagreb, Milan, Rome, Monaco, Nice, Bern and Munich, all within a radius of 600 km.

From the Jožef Stefan Institute, the Technology park “Ljubljana” has been proposed as part of the national strat- egy for technological development to foster synergies be- tween research and industry, to promote joint ventures be- tween university bodies, research institutes and innovative industry, to act as an incubator for high-tech initiatives and to accelerate the development cycle of innovative products.

Part of the Institute was reorganized into several high-tech units supported by and connected within the Technology park at the Jožef Stefan Institute, established as the beginning of a regional Technology park "Ljubljana". The project was developed at a particularly historical moment, characterized by the process of state reorganisation, privatisation and private initiative. The national Technology Park is a shareholding company hosting an independent venture-capital institution.

The promoters and operational entities of the project are the Republic of Slovenia, Ministry of Higher Education, Science and Technology and the Jožef Stefan Institute. The framework of the operation also includes the University of Ljubljana, the National Institute of Chemistry, the Institute for Electronics and Vacuum Technology and the Institute for Materials and Construction Research among others. In addition, the project is supported by the Ministry of the Economy, the National Chamber of Economy and the City of Ljubljana.

Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel.:+386 1 4773 900, Fax.:+386 1 251 93 85
WWW: http://www.ijs.si
E-mail: matjaz.gams@ijs.si
Public relations: Polona Strnad

INFORMATICA

AN INTERNATIONAL JOURNAL OF COMPUTING AND INFORMATICS

INVITATION, COOPERATION

Submissions and Refereeing

Please register as an author and submit a manuscript at: http://www.informatica.si. At least two referees outside the author's country will examine it, and they are invited to make as many remarks as possible from typing errors to global philosoph- ical disagreements. The chosen editor will send the author the obtained reviews. If the paper is accepted, the editor will also send an email to the managing editor. The executive board will inform the author that the paper has been accepted, and the author will send the paper to the managing editor. The paper will be pub- lished within one year of receipt of email with the text in Infor- matica MS Word format or Informatica LaTeX format and figures in .eps format. Style and examples of papers can be obtained from http://www.informatica.si. Opinions, news, calls for conferences, calls for papers, etc. should be sent directly to the managing edi- tor.

SUBSCRIPTION

Please, complete the order form and send it to Dr. Drago Torkar, Informatica, Institut Jožef Stefan, Jamova 39, 1000 Ljubljana, Slovenia. E-mail: drago.torkar@ijs.si

Since 1977, Informatica has been a major Slovenian scientific journal of computing and informatics, including telecommunications, automation and other related areas. In its 16th year (more than twentyeight years ago) it became truly international, although it still remains connected to Central Europe. The basic aim of Informatica is to impose intellectual values (science, engineering) in a distributed organisation.

Informatica is a journal primarily covering intelligent systems in the European computer science, informatics and cognitive community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers ac- cepted by at least two referees outside the author's country. In ad- dition, it contains information about conferences, opinions, criti- cal examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and infor- mation industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the Refereeing Board.

Informatica web edition is free of charge and accessible at http://www.informatica.si.
Informatica print edition is free of charge for major scientific, ed- ucational and governmental institutions. Others should subscribe.

Informatica

An International Journal of Computing and Informatics

Web edition of Informatica may be accessed at: http://www.informatica.si.

Subscription Information Informatica (ISSN 0350-5596) is published four times a year in Spring, Summer, Autumn, and Winter (4 issues per year) by the Slovene Society Informatika, Litostrojska cesta 54, 1000 Ljubljana, Slovenia.
The subscription rate for 2023 (Volume 47) is
– 60 EUR for institutions,
– 30 EUR for individuals, and
– 15 EUR for students
Claims for missing issues will be honored free of charge within six months after the publication date of the issue.

Typesetting: Blaž Mahnič, Gašper Slapničar; gasper.slapnicar@ijs.si
Printing: ABO grafika d.o.o., Ob železnici 16, 1000 Ljubljana.

Orders may be placed by email (drago.torkar@ijs.si), telephone (+386 1 477 3900) or fax (+386 1 251 93 85). The payment should be made to our bank account no.: 02083-0013014662 at NLB d.d., 1520 Ljubljana, Trg republike 2, Slovenija, IBAN no.: SI56020830013014662, SWIFT Code: LJBASI2X.

Informatica is published by Slovene Society Informatika (president Niko Schlamberger) in cooperation with the following societies (and contact persons):
Slovene Society for Pattern Recognition (Vitomir Štruc)
Slovenian Artificial Intelligence Society (Sašo Džeroski)
Cognitive Science Society (Olga Markič)
Slovenian Society of Mathematicians, Physicists and Astronomers (Dragan Mihailović)
Automatic Control Society of Slovenia (Giovanni Godena)
Slovenian Association of Technical and Natural Sciences / Engineering Academy of Slovenia (Mark Pleško)
ACM Slovenia (Ljupčo Todorovski)

Informatica is financially supported by the Slovenian research agency from the Call for co-financing of scientific periodical publications.

Informatica is surveyed by: ACM Digital Library, Citeseer, COBISS, Compendex, Computer & Information Systems Abstracts, Computer Database, Computer Science Index, Current Mathematical Publications, DBLP Computer Science Bibliography, Directory of Open Access Journals, InfoTrac OneFile, Inspec, Linguistic and Language Behaviour Abstracts, Mathematical Reviews, MatSciNet, MatSci on SilverPlatter, Scopus, Zentralblatt Math

Volume 47 Number 3 September 2023 ISSN 0350-5596

Informatica

An International Journal of Computing and Informatics

Informatica **47** (2023) Number 3, pp. 301–468